Suomen konditionaali ja sen vastineita suomi-venäjä -käännöksissä

Korpuspohjainen tutkimus

Petr Halinen
Tampereen yliopisto
Informaatioteknologian ja viestinnän tiedekunta
Monikielisen viestinnän ja käännöstieteen maisteriopinnot
Venäjän kääntämisen ja tulkkauksen opintosuunta
Pro gradu –tutkielma
Toukokuu 2019

Tampereen yliopisto

Informaatioteknologian ja viestinnän tiedekunta

Monikielisen viestinnän ja käännöstieteen maisteriopinnot

Venäjän kääntämisen ja tulkkauksen opintosuunta

Halinen, Petr: Suomen konditionaali ja sen vastineita suomi-venäjä -käännöksissä. Korpuspohjainen tutkimus

Pro gradu -tutkielma, 73 sivua, venäjänkielinen lyhennelmä 12 s.

Toukokuu 2019

Tiivistelmä

Tutkimuksen aiheena on suomen kielen konditionaalin käännösvastineet venäjässä. Tutkimuksella on tarkoitus saada vastaukset seuraaviin tutkimuskysymyksiin: Kääntyykö suomen kielen konditionaali useimmin venäjän kielen konditionaalimuodolla vai jollain muulla tavalla? Mikä tavoista on yleisin, jos ei konditionaali? Vaikuttavatko suomen kielen konditionaalin eri piirteet, kuten aikamuoto tai konditionaalimerkitys käännöstapaan? Vaikuttavatko kääntäjän mieltymykset konditionaalin käännöstapaan?

Sen perusteella, mitä suomen ja venäjän kielten konditionaalimuodoista tiedetään, on tehty seuraava hypoteesi: Konditionaalin päämerkitys on kielissä erilainen, joten oletetaan, että suomen kielen konditionaalimuoto on käännetty suurimmaksi osin jollain muulla tavalla, kuin venäjän konditionaalimuodolla. Yleisintä tapaa kääntää konditionaali ei osata sanoa tässä vaiheessa. Venäjän kielen konditionaali rakentuu useimmiten verbin perfektimuodosta ja suomenkielinen konditionaali voi myös olla aikamuodoltaan perfektissä, joten oletetaan, että perfektiaikamuotoinen konditionaali kääntyy venäjään useammin konditionaalilla, kuin mitä preesensaikamuotoinen.

Tutkimus on toteutettu korpusaineistolla, joka on saatu hakemalla osakorpus Tampereen yliopiston Informaatioteknologian ja viestinnän tiedekunnan Kielten yksikössä laaditusta suomi-venäjä rinnakkaiskorpuksesta ParFin. Osakorpus on rajattu teoksiin, jotka on julkaistu vuoden 1976 jälkeen ja valittu osakorpus sisältää 12 kirjailijan 16 eri teosta. Varsinainen aineisto on osakorpuksesta haetut 1000 konditionaaliesimerkkiä ja niiden käännösvastineet. Aineiston analysointimenetelminä on käytetty binääristä logistista regressiota, ehdollisen päättelyn puuta (conditional inference tree) ja satunnaista metsää (random forest).

Tutkimuksen tuloksena hypoteesin molemmat oletukset vahvistuivat. Suomen kielen konditionaalimuoto on käännetty suurimmaksi osin muilla tavoin kuin venäjän konditionaalilla, jolla on käännetty 28,6 % esimerkeistä. Yleisin venäjänkielinen käännösvastine on *aikamuodoltaan futuurissa oleva perfektiaspektinen verbi, jonka tapaluokka on indikatiivi*. Perfektiaikamuotoisista konditionaalimuodoista 44 % kääntyy konditionaalilla venäjään, kun taas preesensaikamuotoisista esimerkeistä vain 24 % käännettiin konditionaalilla.

Satunnaisen metsän tulosten perusteella kääntäjä vaikuttaa konditionaalin käyttöön käännösratkaisuna lähes puolessa (46,9 %) tapauksista. Konditionaalimerkitys vaikuttaa lähes kolmasosassa (31,7 %) ja konditionaalin aikamuoto noin viidesosassa (21,6 %) tapauksista.

Avainsanat: konditionaali, korpukset, kääntäminen, binäärinen logistinen regressio.

SISÄLLYS

1. JOHDANTO

Konditionaali eroaa muista modaalisuuden ilmaisukeinoista siten, että siinä mahdollisuuden merkitys esiintyy eri tavalla. Konditionaali osoittaa suunniteltua, kuviteltua tai ennustettua asianlaitaa. Siinä suunnitellut asiat ovat olemassa ajattelun, mielikuvituksen, päättelyn ja tahdon tasolla, mutta eivät vielä oikeassa maailmassa. Tällaisessa tilanteessa on kyse intensionaalisesta asiaintilasta eli mahdollisuudesta yhtenä ajateltavissa olevana vaihtoehtona. (VISK - § 1592.)

1.1 Tutkimuksen tausta ja tavoite

Huomasin jo ala-asteella ollessani, että tietyt ihmiset olivat lahjakkaita eri asioissa. Tarkemmin, toiset ovat matemaattisesti lahjakkaita ja toiset, kuten itse koen olevani, ovat kielellisesti lahjakkaita. Huomasin nimenomaan eräänlaisen korrelaation näiden välillä. Matemaattisesti lahjakkaat eivät useinkaan olleet lahjakkaita kielellisesti ja päinvastoin en esimerkiksi itse ns. kielellisesti lahjakkaana osannut enkä myöskään ollut kiinnostunut matematiikasta. En tiedä onko tämä totta vai ei, mutta tämmöisiä havaintoja tein ainakin itsestäni ja koulukavereistani jo ala-asteella.

Kasvoin kaksikielisessä perheessä ja ympäristössä. Kotona puhuttiin aina lähinnä venäjää ja kotini ulkopuolella luonnollisesti suomea. Kun olin pieni, ei äitini osannut suomen kieltä kovinkaan hyvin, joten yrittäessäni selittää äidilleni koulussa tehtyjä asioita, täytyi minun joko yrittää kuvailla asiat venäjäksi tai pyytää suomen kieltä hyvin osaavaa isääni kääntämään ne puolestani, jolloin opin asioiden venäjänkielisiä nimityksiä. Tästä syystä uskon myös, että olen ollut aina kiinnostunut kielten erilaisista piirteistä ja niiden käännöstavoista sekä ylipäätään kääntämisestä. Minua ovat myös aina kiinnostaneet esimerkiksi eri sanojen syntyperät. Halusin siis valita tutkielmaani varten sellaisen aiheen, joka kiinnostaisi minua, mutta josta ei myöskään löytyisi suurta määrää edeltäviä tutkimuksia. Päädyin tätä kautta tutkimaan suomen konditionaalin kääntämistä venäjän kielelle.

Lähtökohtaisena tutkimusongelmana on se, että konditionaalia on tutkittu paljon, mutta sen kääntämistä ja käännösvastineita suomi-venäjä -kieliparissa ei. Halusin siis selvittää, millä mahdollisilla eri tavoilla suomen kielen konditionaali kääntyy venäjän kieleen. Eniten minua kiinnosti tietää aidon kielenkäytön korpuspohjaisen materiaalin perusteella, että käytetäänkö suomen kielen konditionaalimuodon kääntämiseen eniten venäjän kielen konditionaalia vai jotain muuta käännösratkaisua. Lisäksi halusin selvittää, mikä on yleisin tapa kääntää suomen kielen konditionaali ja mitkä asiat vaikuttavat siihen, että käytetäänkö käännösratkaisuna konditionaalia.

Tutkimukseni tavoitteena on selvittää vastauksia seuraaviin kysymyksiin:

1. Kääntyykö suomen kielen konditionaali useimmin venäjän kielen konditionaalimuodolla vai jollain toisella tavalla?
2. Mikä tavoista on yleisin, jos ei konditionaali?
3. Vaikuttavatko suomen kielen konditionaalin konditionaalimerkitys, aikamuoto tai semanttinen merkitys käännöstapaan?
4. Vaikuttavatko kääntäjän mieltymykset käännöstapaan?

Yhtenä tutkimustavoitteistani voisi sanoa olleen myös konditionaalin kääntämisen tutkiminen korpusaineiston avulla suomi-venäjä -kieliparissa. Toivon lisäksi, että tutkimuksestani voisi olla tulevaisuudessa hyötyä jollekin toiselle tämän aiheen tutkijalle. Mielestäni tätä on mielenkiintoista ja kannattavaa tutkia, koska aihettani ei ole paljoakaan tutkittu ja tutkimustuloksista saadaan uutta tietoa konditionaalin kääntämisestä ja käytöstä suomi-venäjä -käännöksissä.

1.2 Aiheeseen liittyvä aiempi tutkimus

Niin venäjän kuin suomenkin kielen konditionaalia on tutkittu yleisellä tasolla hyvin paljon. Esimerkiksi erilaisia kielioppeja ja tutkimuksia, joissa käsitellään konditionaalia, on paljon.

Kuitenkin etsiessäni tutkimuksia, joissa tutkittaisiin konditionaalin kääntämistä näiden kielten välillä, löysin ainoastaan yhden tutkimuksen, joka käsitteli omaa tutkimusaihettani: Kirsi Kuuselan Tampereen yliopistolla tekemä pro gradu -tutkielma vuodelta 1992 nimeltä ”Soslagatel'noe naklonenie v sovremennom russkom jazyke i ego perevod na finskij jazyk : Konditionaali nykyvenäjässä ja sen kääntäminen suomen kielelle”. Minun tutkielmani käsittelee kuitenkin konditionaalin kääntämistä juuri päinvastaisesti eli suomesta venäjään. Tutkielmani on lisäksi korpuspohjainen, jossa on käytetty materiaalina aitoja kaunokirjallisia tekstejä ja niiden käännöksiä.

1.3 Teoria & Tutkimusaineisto

Tutkimukseni on korpusaineistolla toteutettu, joten tästä syystä osa tutkielman teoriaosuudesta käsittelee korpuksia ja niiden eri käyttötapoja lingvistisissä ja käännöstieteellisissä tutkimuksissa. Teoriaosuudessa käsittelen lisäksi tietenkin myös konditionaalia ja vertailen suomen ja venäjän konditionaalimuotoja keskenään.

Varsinainen tutkimusaineistoni koostuu 1000 suomenkielisen konditionaalin käytön esimerkistä ja niiden venäjänkielisistä käännösvastineista. Aineisto haettiin Tampereen yliopiston käännöstieteen

tutkijoiden laatimasta kaunokirjallisuuden rinnakkaiskorpuksesta *ParFin* (suomi-venäjä). Korpuksesta haettiin osakorpus, johon kuuluvat tekstit, jotka on julkaistu vuoden 1976 jälkeen. Tämä rajaus tehtiin, koska ei haluttu aineistoon konditionaalin käytön esimerkkejä erittäin vanhoista teoksista, kuten esimerkiksi ParFin korpuksessa oleva Juhani Ahon teos Rautatie vuodelta 1884. Osakorpuksesta haettiin tämän jälkeen paralleelikonkordanssihaulla 1000 suomenkielistä konditionaaliesimerkkiä satunnaisessa järjestyksessä. Haun tuloksena saatiin 1000 konditionaaliesimerkkiä ja niiden rinnastetut käännösvastineet. Molemmat näistä olivat hakutuloksessa omissa konteksteissaan.

1.4 Tutkimusmetodien esittely

Tutkimusmetodini ovat *binäärinen logistinen regressio* ja jatkotoimenpiteet *ehdollisen päättelyn puu* (conditional inference tree) ja *satunnainen metsä* (random forest). Nämä toimenpiteet suoritettiin ohjelmalla R, josta lisää luvussa 4.4.

Binäärisellä logistisella regressiolla saadaan tieto siitä, mitkä tekijät (selittävät muuttujat) vaikuttavat tutkittavaan tulokseen (vastemuuttujaan) huomattavalla tavalla, kuinka suuresti ne vaikuttavat siihen ja mihin suuntaan ne vievät saatua tulosta (De Sutter 2018).

Ehdollisen päättelyn puu on metodi, joka perustuu kaksijakoiseen toistuvaan ositukseen (Merkle & Shaffer 2011). Tulokseksi saadaan puukuvio, jossa näkyy eri tekijöiden vaikutus toisiinsa ja tutkittavaan muuttujaan.

Satunnaisessa metsässä luodaan puumetsä (esim. 500 ehdollisen päättelyn puuta), joka on tehty selittävien muuttujien ja eri datapisteiden satunnaisella valinnalla. Metsän puiden tulokset sulautetaan yhteen ja sen arvot kertovat eri selittävien muuttujien vaikutuksista vastemuuttujaan muiden selittävien muuttujien yhteydessä. Tuloksena saadaan tarkka käsitys siitä, mitkä selittävät muuttujat vaikuttavat vastemuuttujaan eniten. (Levshina 2015.)

Lisää metodeista luvussa 4.2.2.

1.5 Aineiston käsittely

Käsittelin korpushaun avulla saadun aineiston aluksi Microsoft Excelissä. Korpushaulla saadussa aineistossa on valmiiksi metadataa konditionaaliesimerkeistä ja käännösvastineista, kuten kirjailija, teoksen suomenkielinen nimi, kääntäjä ja teoksen venäjänkielinen nimi. Näiden lisäksi loin varsinaisia käsittelysarakkeita, joissa luokittelin esimerkit niiden aikamuodon,

konditionaalimerkityksen ja verbin semanttisen merkityksen mukaan. Toimin samoin käännösvastineiden osalta, jotka luokittelin sanaluokan, tapaluokan, aikamuodon ja aspektin mukaan sekä loin vielä sarakkeen, johon on merkattuna, että onko käännösvastine konditionaalissa.

Tämän jälkeen siirryin käsittelemään aineistoa ohjelmalla R, jonka avulla pystyin tekemään aineistoni analyysin valitsemillani metodeilla. Tarkoituksenani oli käyttää binääristä logistista regressioanalyysiä (luku 4.4) ja jatkotoimenpiteitä ehdollisen päättelyn puut (luku 4.5.1) sekä satunnaiset metsät (luku 4.5.2) aineistoni kaksijakoisen vastemuuttujan analysoimiseen. Ja selvittää siten, mitkä seikat (luomani selittävät muuttujat) vaikuttavat siihen, että suomen kielen konditionaalimuoto käännetään venäjään konditionaalilla.

1.6 Tutkielman rakenne

Ensimmäisessä luvussa käyn läpi tutkimuksen taustan ja tavoitteen, avaan tutkimusaineistoa ja tutkimusmetodeja, kerron aineiston käsittelystä ja käyn läpi tutkielman rakenteen.

Toisessa luvussa kerron korpuksista yleisesti, korpuslingvistisestä tutkimuksesta ja korpusvetoisesta käännöstieteellisestä tutkimuksesta.

Kolmannessa luvussa tarkastelen konditionaalia suomessa ja venäjässä. Kerron mitä niistä tiedetään ja vertailen niitä keskenään. Tässä luvussa teen myös hypoteesini teorian pohjalta.

Neljännessä luvussa tarkastelen suomen konditionaalin vastineita korpusaineistossa. Kerron käyttämästäni korpuksesta ja aineiston hausta. Avaan tutkimusmetodiani ja aineistoani tarkemmin eli kerron aineistoni tilastot. Analysoin aineistoni käyttäen valitsemiani tutkimusmetodeja ja kerron luvun lopussa vahvistuiko hypoteesini vai ei.

Viidennessä luvussa pohdin, milloin konditionaalia tarvitaan venäjään kääntäessä. Teen yhteenvedon tutkielmastani ja käyn läpi muuta pohdintaa sekä mahdollisia jatkotutkimusaiheita.

2. KORPUSVETOINEN LINGVISTINEN JA KÄÄNNÖSTIETEELLINEN TUTKIMUS

2.1 Korpuksista yleisesti

Sähköisten tekstien yleistymisen myötä, mukaan lukien lähes loputtoman internetin tarjonnan, on paljon mahdollisuuksia sovittaa yhteen erikielisiä tekstejä erilaisia kieltenvälisiä tutkimuksia varten. Internetiä itseään voidaan pitää valtavana korpuksena, jota voidaan käyttää, jos muuta sopivaa tekstikorpusta ei löydy. Kaksikielisiä korpuksia on käytetty hyödyllisinä resursseina mm. *translationese* ilmiön tutkimuksessa eli tutkittaessa tapoja, joilla kääntäjään vaikuttavat alkuperäiskielen rakenteet ja ilmaisut. Tästä lisää Johanssonin (2007) kirjassa.

Yksi määritelmä korpukselle on: tekstikokoelma, joka on koottu tiettyä tarkoitusta varten. Toisin sanottuna, korpus ei ole vain kokoelma satunnaisia tekstejä, vaan se on tekstikokoelma, joka on koottu tiettyjen kriteerien mukaan. Yksi korpuksen kriteereistä on se, että sen on tarkoitus olla, jotain edustava. (Cheng 2012, 3.) Esimerkiksi *ParFin* korpus edustaa kaunokirjallisia tekstejä ja niiden käännöksiä.

Lingvistinen tai kielellinen tekstikorpus ymmärretään yleensä isona, koneluettavassa muodossa olevana, standardisoituna, strukturoituna ja annotoituna massiivisena kielitiedostona, joka on suunniteltu konkreettisten lingvististen ongelmien ratkaisuun (Zaharov & Bogdanova 2011, 7). Tämä ei kuitenkaan tarkoita, että korpuksia käyttävät ainoastaan korpuslingvistit (Cheng 2012, 3). Korpusta voidaan käyttää muillakin aloilla, kuten kirjallisuustieteessä, käännösteorian tutkimuksessa jne.

Suurin osa korpuksista, jos niiden kokoaminen ei ole kesken, sisältävät kiinteän määrän sanoja. Esimerkiksi BNC -korpus (British National Corpus) - 100 miljoonaa sanaa tai Cobuild Bank of English -korpus - 650 miljoonaa sanaa. On silti olemassa myös korpuksia, joita päivitetään ja laajennetaan koko ajan, kuten Tampereen yliopiston ParFin tai ParRus korpukset. Yksi syy päivittämiselle on kielen muutosten seuraaminen. Toinen syy on se, että korpuksen kasaaminen on aikaa vievää, minkä takia korpuksen kokoaminen pitää suunnitella siten, että siitä hakeminen on mahdollista ennen kuin kaikki tekstityypit, aiheet ja aikakaudet ovat täysin edustettuina. Tämmöisistä korpuksista tutkitaan koko korpuksen sijaan useimmiten jotain korpuksen osaa eli sen osakorpusta. Näin teen myös tässä tutkielmassa. Osakorpukset ovat tekstijoukkoja, jotka korpuksen tekijä tai käyttäjä määrittelee eli saman genren tekstit, saman kirjoittajan tekstit, tietyn aikakauden tekstit jne.

ja tällä tavoin osakorpusta käyttämällä voi isosta avoimesta korpuksesta saada käyttöönsä kiinteämääräisen korpuksen. (Mikhailov & Cooper 2016, 7-8.)

Monikielisten korpusten tutkimus kehittyy koko ajan, joten myös erilaisten tutkimusten määrä kasvaa jatkuvasti. Monikielistä korpusta on käytetty ainakin seuraavissa tutkimuksissa: kontrastiivinen lingvistiikka, kielen typologia, käännöstiede ja kääntäjien koulutus, kaksikielinen leksikografia, vieraan kielen opetus ja luonnollisen kielen prosessointi (ml. konekääntäminen). Uusien tutkimusmahdollisuuksien syntyminen näyttäisi olevan monikielisille korpuksille ominaista. Iso osa tutkimuksista on deskriptiivistä ja soveltavaa (jonkun tietyn käytännöllisen ongelman ratkaisua), mutta monikielisiä korpuksia käytetään myös lingvistisen teorian testaamiseen ja kehittämiseen. Vaikka tutkimusala kehittyy jatkuvasti, niin siinä on silti ongelmia, joista osa liittyy korpuksen kokoamiseen. Yksikään yksittäinen korpus ei ole riittävä kaikkiin tarkoituksiin. Eri korpuksia ja korpusmalleja tarvitaan tutkimuskysymyksestä ja tutkimuksen tarkoituksesta riippuen. Tietyt korpukset on tehty erityisesti jotain tutkimusta varten, kuten kaksisuuntainen käännöskorpus, joka on tehty kontrastiivista tutkimusta varten. (Johansson 2007, 301.)

Erilaisia korpuksia on paljon, mutta niitä voidaan lajitella erilaisten kriteerien mukaan. Kielellisten tietojen perusteella lajiteltuna korpukset voi jakaa teksti-, puhe- ja sekakorpuksiin. Sekakorpukset sisältävät tietenkin sekä tekstiä että puhetta. Korpukset voidaan luokitella myös rinnakkaisuuden perusteella: yksikielisiin, kaksikielisiin tai monikielisiin korpuksiin. Kirjakielisyyden kriteerien mukaan korpukset jaotellaan kirjakielisiin, murteellisiin, puhekielisiin, terminologisiin ja sekakorpuksiin. Korpukset voidaan myös lajitella tavoitteen mukaan: useita tavoitteita varten koottuihin ja yhtä erityistavoitetta varten koottuihin korpuksiin. (Zaharov & Bogdanova 2011, 22-23.)

Tämän tutkielman tutkimuksessa käytetty korpus *ParFin* (Parallel Finnish-Russian Corpus) on näillä eri kriteereillä jaoteltuna: tekstikorpus, kaksikielinen yhdensuuntainen rinnakkaiskorpus (suomi-venäjä), kirjakielisyyden mukaan kirjakielinen (koostuu kaunokirjallisista teoksista) ja se on useita eri tavoitteita varten luotu.

2.1.1 Rinnakkaiskorpus (käännöskorpus)

Sekaannuksen välttämiseksi mainittakoon, että tässä tutkielmassa päätin käyttää termejä rinnakkaiskorpus ja käännöskorpus tarkoittamaan samaa asiaa. Mainitaan lisäksi Mikhailovin & Cooperin (2016) määritelmä käännöskorpukselle: korpus, joka koostuu tietylle kielelle käännetyistä teksteistä. Se voi sisältää myös saman kielisiä alkuperäistekstejä vertailua varten, jolloin käännetyt

tekstit eivät ole korpuksen alkuperäiskielisten tekstien käännöksiä. Heidän mukaansa tämmöisistä korpuksista on hyötyä, kun arvioidaan yhden tai useamman käännöksen kieltä vertailemalla niitä kieleen, jota käytetään alkuperäiskielisissä teksteissä. (Mikhailov & Cooper 2016, 220.)

Johanssonin (2007, 9) määritelmä termille *käännöskorpus* (korpus, jossa on alkuperäistekstejä ja niiden käännöksiä yhdellä tai useammalla kielellä) vastaa Mikhailovin & Cooperin (2016, 219) määritelmää termille *rinnakkaiskorpus* (kaksikielinen tai monikielinen korpus, jossa on alkuperäistekstejä ja niiden käännöksiä yhdellä tai useammalla kielellä) lähes täydellisesti. Eli tässä tutkielmassa puhuttaessa käännöskorpuksesta tai rinnakkaiskorpuksesta tarkoitetaan samaa asiaa.

Rinnakkaiskorpukset ovat siis kaksikielisiä tai monikielisiä korpuksia, jotka koostuvat alkuperäisteksteistä ja niiden käännöksistä yhdelle tai useammalle kielelle (Mikhailov & Cooper 2016, 5). Rinnakkaiskorpus on tavallisesti merkattu, jotta hakujen tekeminen helpottuu. Merkkaaminen tarkoittaa sitä, että sanojen epämääräiset piirteet merkataan erityisillä tageilla. Yleisin tapa merkata on *lemmatisaatio* eli *annotaatio*, joka osoittaa joka sanan perusmuodon (esim. *Ottaa*, muodoille *otan, otat, ottaa, otatte, otamme ja ottavat*). Lemmatisaatio yhdistetään yleensä *tageilla* merkitsemiseen, jossa merkitään sanaluokka ja joissain tapauksissa siihen liitetään myös morfologista informaatiota (akkusatiivi, genetiivi; konditionaali, perfekti, jne.) (Mikhailov & Cooper 2016, 3.) Morfologisesti annotoidusta korpuksesta on mahdollista löytää tietoa tiettyjen kielioppimuotojen yleisyydestä ja käytöstä, kuten esimerkiksi konditionaali, jota tutkin tässä tutkielmassa.

Käännöskorpuksista tutkitaan *ekvivalensseja*. Niistä puhuttaessa tarkoitetaan vastaavuuksia, joita käytetään sitten todisteena samankaltaisuudesta tai eroavaisuudesta tai käännösprosessin luomasta ominaisuudesta. Käännösprosessin luomia erikoispiirteitä tai ominaisuuksia kutsutaan myös *käännösefekteiksi*. (Johansson 2007, 5.)

Suurin osa käännöskorpuksista eli rinnakkaiskorpuksista on yhdensuuntaisia (suomi-venäjä), mutta on myös kaksisuuntaisia käännöskorpuksia (venäjä-suomi-venäjä). Kaksisuuntaisen käännöskorpuksen avulla voi paikantaa alkuperäistekstien ja saman kielisten käännösten kielimuotovalintojen eroavaisuudet (kontrastiivinen analyysi). Joissain tapauksissa vastaan tulee ylikäyttöä eli kielimuotoa on käytetty käännöksissä useammin kuin saman kielisissä alkuperäisteksteissä. Toisissa tapauksissa vastaan tulee alikäyttöä, jolloin frekvenssi-ero on käänteinen. Ylikäyttö ja alikäyttö voidaan nähdä todisteena sille, että ilmaisun keinot eivät vastaa toisiaan lähde- ja kohdekielissä, ja että lähdetekstillä on taipumus jättää jälkensä käännökseen. (Johansson 2007, 32.)

Käännöskorpuksen hyviä puolia: se sisältää tekstejä, joiden tarkoituksena on ollut ilmaista sama merkitys ja niillä on sama diskurssifunktio molemmissa kielissä. Tekstit ovat lisäksi yleensä suosittuja ja paljon luettuja. (Johansson 2007, 9-10.)

Käännöskorpuksen ongelmia: käännettyjen tekstien määrä on usein hyvin rajattu, verrattuna alkuperäiskielisten tekstien määrään. Tekstien määrä ja laji saattavat myös vaihdella riippuen käännössuunnasta. Teksteissä käytetyt valinnat vaihtelevat myös yksittäisten kääntäjien mukaan tai saattavat olla jopa virheellisiä. Mistä voi siis käännöstekstien perusteella tietää, onko molemmilla kielillä ilmaistu sama asia tai merkitys tai miten hyvin käännösteksti vastaa oikeaa tavallista kielenkäyttöä? Näitä käännösefektejä pystytään käännöskorpusta käytettäessä hallitsemaan, jos käytetään samanaikaisesti käännös- ja vertailukorpusta. Tästä syystä johtuen on tärkeää yhdistää käännöskorpus ja vertailukorpus. (Johansson 2007, 9-10.)

2.1.2 Vertailukorpus

Alkuperäiskielisiä tekstejä sisältäviä monikielisiä korpuksia voidaan kutsua vertailukorpuksiksi. Vertailukorpukset ovat kokoamisen kannalta vähemmän ongelmallisia, kuin käännöskorpukset. Jos saatavilla on yksikielisiä korpuksia eri kielillä, niin on mahdollista käyttää niitä vertailtavien tekstien yhteensovittamisen lähteenä. (Johansson 2007, 303.)

Mikhailovin & Cooperin (2016) tarkka määritelmä vertailukorpukselle on korpus, joka koostuu kahdesta tai useammasta tekstikokoelmasta, jotka on koottu samojen kriteerien (kokoelmien koko, esimerkkien koko, samat aiheet, sama aikakausi jne.) mukaan. Ne voivat koostua eri kielten tai saman kielen eri variaatioiden (esim. ruotsin kieli Suomessa ja Ruotsissa) teksteistä. (Mikhailov & Cooper 2016, 217.)

Vertailukorpuksen hyviä puolia: se sisältää alkuperäiskielisiä tekstejä, jotka kuvastavat tavallista kielenkäyttöä eri kielissä ja joiden avulla voi turvallisesti tehdä johtopäätöksiä vertailussa olevien kieltenvälisistä eroista ja samankaltaisuuksista (Johansson 2007, 10). Vertailukorpusten käyttäjät voivat lisäksi käyttää resursseja ja työkaluja, jotka on kehitetty yksittäisiä kieliä varten tai yksikielisiä korpuksia varten.

Vertailukorpuksen ongelmia: vertailussa huomatut erot saattavat johtua siitä, että kieltenvälisessä rinnakkaistamisessa on tapahtunut virhe. Vaikeinta on tietää, mitä verrata eli millä kielimuodoilla on sama merkitys ja käytännöllinen funktio molemmissa vertailtavissa kielissä. (Johansson 2007, 10.)

Molempia näistä korpusmalleista (rinnakkaiskorpus, vertailukorpus) käytetään vertailututkimukseen, jota tehdään esim. leksikologiassa, kielioppien tekemisen yhteydessä ja käännöstieteessä sekä käännösmetodien kehittämiseen (mm. konekäännös) (Zaharov & Bogdanova 2011, 26). Patrizia Giamperi (2018, 239) toteaakin, että vertailukorpus ja rinnakkaiskorpus ovat erittäin hyödyllisiä yhdessä käytettynä, koska rinnakaiskorpus antaa tietoa aiemmin käytetyistä käännösratkaisuista ja vertailukorpus antaa esimerkkejä näiden käännösratkaisujen aidosta käytöstä.

Nämä kaksi korpustyyppiä täydentävät siis toisiaan ja antavat mahdollisuuden kontrolloida käännösefektejä: mitkä käännöksen perusteella paljastuneet kieltenväliset vastaavuudet ovat kielen oikeita edustajia, ja mitkä taas ovat käännöksen tulos.

2.1.3 Korpusten käyttö tutkimuksissa

Korpusten käytön voi perustella seuraavilla asioilla. Korpuksen tarpeeksi suuri representaatio (koko) takaa siitä saatujen tietojen luotettavuuden ja tarjoaa kattavan kokonaiskuvan kaikista kielellisistä ilmiöistä. Korpuksesta löytyvät erilaiset tiedot esiintyvät siinä omassa aidossa kontekstissaan, mikä tarjoaa mahdollisuuden tutkia niitä kaikinpuolisesti ja objektiivisesti. Kerran koottu ja valmisteltu korpus voi olla käytössä useita kertoja eri tutkijoiden toimesta ja eri tarkoituksia varten. (Zaharov & Bogdanova 2011, 8.)

Korpuksen käyttö voidaan nähdä ikään kuin sellaisena dialogina tutkijan ja materiaalin välillä, jossa yhden kysymyksen tutkiminen saattaa johtaa lisäkysymyksiin. Tavallisesti tutkimukset seuraavat tiettyä kaavaa: tutkimuskysymysten muotoilu, korpuksen tutkiminen, toistuvien kaavojen löytäminen ja datan tulkinta. Mistä tutkimuskysymykset syntyvät? Joskus aiemman tutkimuksen perusteella, joskus huomioiden tai intuition perusteella. Joissain tapauksissa motivaattoreina ovat metodiikan kehittäminen ja sen testaus. Korpuksen tutkiminen ilman tutkimuskysymystä on kuin etsisi neulaa heinäsuovasta. (Johansson 2007, 38-39.)

Korpukset ovatkin usein eri tutkijoiden käytössä useita eri tarkoituksia varten. Korpusten eri käyttäjiä ovat esim. lingvistiteoreetikot, jotka käyttävät korpuksia kokeilupohjana omien hypoteesiensa tarkistamiseen ja oikeaksi todistamiseen. Soveltavat lingvistit (opettaja, kääntäjä yms.) käyttävät sähköisiä korpuksia kielenoppimisessa tai -opetuksessa ja muiden ammatillisten ongelmien ratkaisussa. Kääntäjien tapauksessa luonnollisin käyttötapa on käännösongelmien ratkaisu. Korpusten erityinen käyttäjäryhmä ovat kieliteknologian tutkijat, jotka yrittävät kartoittaa ja käyttää teksteissä olevia tilastollisia ja lingvistisiä sääntöjä, ja luoda niiden perusteella kielen tietokonemallinnoksen. (Zaharov & Bogdanova 2011, 94-95.)

Korpusten käyttö hypoteesien vahvistamiseksi, kumoamiseksi tai jonkun uuden ja odottamattoman näkemyksen luomiseksi on muuttanut sen, miten kieltä tutkitaan. Nykyään monet kieliopit ja sanakirjat perustuvat korpusdataan, vaikka joissakin ainoastaan kielenkäyttöesimerkkien osalta. Ensimmäisen kokonaisvaltaisen kontrastiivisen kieliopin tuottaminen on kuitenkin vain unelma. Tämä johtuu siitä, että sellaisen tuottaminen tapahtuu pikkuhiljaa ja vaatii monikielisestä korpuksesta saatujen todisteiden tutkimista monien vuosikymmenten ajan. Tämmöinen korpus voisi kuitenkin antaa vastauksia ongelmiin, joita esiintyy käännöstieteessä, kontrastiivisessa lingvistiikassa, kielenopetuksessa ja kielen universaalien opiskelussa. (Mikhailov & Cooper 2016, 15-16.)

Monien käyttäjien lisäksi korpuksilla on siis monia käyttötapoja, mutta niistä yleisimpiä ovat empiirinen tuki, frekvenssitieto ja metainformaatio. Monet lingvistit (kuten lingvistiteoreetikot) käyttävät korpusta ns. esimerkkipankkina eli yrittävät löytää korpuksesta empiiristä tukea hypoteeseille ja säännöille, joiden parissa he tekevät tutkimustaan. Empiirinen tuki on laadullinen (kvalitatiivinen) tutkimusmetodi korpuksen tutkimukseen. Korpuksesta voidaan kuitenkin saada myös tietoa sanojen, fraasien ja rakenteiden frekvensseistä. Tätä tietoa voidaan käyttää määrälliseen (kvantitatiiviseen) tutkimukseen. Kielellisen sisällön lisäksi korpukset sisältävät paljon ekstralingvististä informaatiota eli metainformaatiota. Metainformaatio tarkoittaa sellaista tietoa, kuten kirjoittajan tai puhujan ikä ja sukupuoli, tekstin genre, tekstin alkuperä ja aikakausi jolta teksti on jne. (Zaharov & Bogdanova 2011, 94-97.)

Korpuksen käytöllä on myös omia ongelmiaan. Korpuksesta saatu data voi olla luonteeltaan joko hyvin yleistasoista tai hyvin tarkasti määriteltyä tietoa. Yleistasoinen data pitää sisällään kokonaisia tekstejä, tekstiryhmiä tai jopa kokonaisen korpuksen. Data voi olla frekvenssilistojen, kollokaatiolistojen tai tekstilistojen muodossa eli esimerkiksi korpuksen sanamäärä, keskimääräinen sanojen pituus jne. Tarkasti määritelty data tarkoittaa oikeita kielenkäytön esimerkkejä, joita korpuksesta saadaan. Asiat, joita haetaan ja niitä ympäröivä konteksti ovat yleensä melko lyhyitä (rivi, virke tai kappale), ja ne saadaan konkordanssien muodossa. Molemmista näistä on hyötyä tutkimuksissa. Yleistasoinen data antaa tutkijalle yleiskuvan tilastojen kautta ja tarkasti määritelty esimerkkien muodossa oleva data antaa todisteet ja perustelut yleistason datan tilastoille. (Mikhailov & Cooper 2016, 8-9.)

2.1.4 Korpushaku

Näille kaikille käyttötavoille ja käyttäjille yhteistä on se, että mikä tahansa informaatio täytyy hakea korpuksesta. Haku tapahtuu yleensä jollain korpushallintasovelluksella. Korpushallintasovellus on ohjelma, jonka avulla voi hakea korpuksesta haluamiaan asioita. Korpushaulla voidaan hakea mille

tahansa sanalle konkordanssi – lista kaikista sanan esiintymisistä kontekstissa ja linkki lähteeseen. (Zaharov & Bogdanova 2011, 8.)

Nykyään lähes kaikilla netistä löytyvillä korpuksilla on oma nettisovellus eli aiemmin mainittu korpushallintasovellus, joka on yleensä konkordanssihakemiston muodossa. Konkordanssihakemisto on hakukone, joka tuottaa tuloksen konkordanssin muodossa. Joskus ne sisältävät muitakin työkaluja, joilla voi laskea frekvenssejä tai tarkastella korpuksen tilastoja. Korpuksen teksteihin pääsee käsiksi vain hakukäyttöliittymän kautta. Teksteihin ei ole mahdollista päästä muuten käsiksi, eikä niitä voi ladata omalle koneelleen. Joissain tapauksissa korpukset ovat ladattavissa tekstitiedostohakemistoina, annotaatio-informaation sisältävinä XML tiedostoina tai kohdistettuina teksteinä, käännösmuistiohjelmien käyttämässä TMX formaatissa. Ladattavissa olevat korpukset tai tekstit ovat yleensä tekijänoikeusvapaita tekstejä. (Mikhailov & Cooper 2016, 45.)

Useimman korpuspohjaisen tutkimuksen alkupisteenä on yleensä frekvenssilistojen haku. Frekvenssilista on lista tekstin tai tekstien sanoista ja niiden yleisyydestä, ja se näyttää minkälaista dataa korpuksesta on mahdollista löytää. Frekvenssilistan katsominen on myös hyvä tapa varmistaa, että korpus sisältää tarpeeksi paljon sopivaa tutkimustietoa. Sillä, jos korpuksesta on mahdollista saada vain pieni näyte jostain asiasta, ei ole mahdollista käyttää sitä todisteena asian yleisistä piirteistä. (Mikhailov & Cooper 2016, 51-52.)

Kuten aiemmin mainittiin, rinnakkaiskorpuksista haetaan ekvivalensseja tai vastaavuuksia (luku 2.1.1). Tavallisilla hakumenetelmillä on mahdollista saada kahdenlaisia konkordansseja: yksikielisiä tai rinnakkaiskonkordansseja (paralleeli). Rinnakkaiskonkordansseista vertaillaan käännösvastineita. Niitä käytetään yleisimmin kahteen tarkoitukseen: halutaan selvittää joku yksinkertainen käytännöllinen kysymys esim. miten tietty sana kääntyy kielestä A kieleen B. Tai halutaan dataa, josta voi tutkia jotain laajempaa ongelmaa, kuten esim. toisiaan vastaavien kielirakenteiden suhteet eri kielissä. (Mikhailov & Cooper 2016, 60.)

2.1.5 Hakutulosten käsittely

Korpuspohjaista lingvististä tutkimusta tehdessä huomaa nopeasti, että ensimmäiset hakutulokset eivät useimmiten tarjoa lopullista vastausta tutkimuskysymykseen. Yleensä vaaditaan hakutulosten käsittelyä ja lisäanalyysiä, joten usein on järkevintä ladata saadut hakutulokset ohjelmaan, joka on suunniteltu erityisesti suurien datamäärien käsittelyyn. Yksi helpoimmista tavoista, ja tapa, jota käytettiin ladattujen hakutulosten ensikäsittelyyn tässäkin tutkielmassa, on käyttää taulukkolaskentaohjelmaa esim. Microsoft Excel. Muutkin samankaltaiset taulukko-ohjelmat käyvät

hyvin, mutta erittäin suurten tietomäärien käsittelyyn kannattaa käyttää tietokantaohjelmistoja (esim. Microsoft Access tai LibreOffice Base) tai tilasto-ohjelmistoja (esim. SPSS tai R).

On olemassa monenlaisia hakutuloksena saatuja konkordanssitaulukoita. Joissain voi olla valmiiksi monta saraketta, joihin ovat merkittynä esim. julkaisuvuosi, kirjoittajan ja kääntäjän nimet, alkuperäisen teoksen ja käännöksen nimet. Näin oli esim. tämän tutkielman materiaalin osalta. Tämä riippuu kuitenkin siitä, mitä konkordanssihakuohjelma antaa tulokseksi ja missä muodossa saatu data on. Hakutulosten käsittelyssä tehdyt eri toimenpiteet riippuvat muutenkin pitkälti siitä, mitä yritetään saada aikaiseksi, missä muodossa tieto on ja mitä tutkija itse haluaa tehdä. (Mikhailov & Cooper 2016, 93-94.)

Täytyy kuitenkin muistaa, että korpuksesta saadut tulokset eivät ole välttämättä täysin luotettavia, koska niissä saattaa näkyä kääntäjien mieltymykset eikä ns. objektiivinen totuus. On kuitenkin mahdollista tarkistaa, ovatko korpuksesta saadut tulokset kääntäjien mieltymyksistä johtuvia vai eivät. On olemassa Chi-square testi. Chi-square testejä käytetään havainnoidun datan itsenäisyyden tarkistamiseen (eli onko se luotettavaa vai sattumaa). Testiä varten havainnoidut arvot tallennetaan taulukkoon, jossa toisiaan vastaavat arvot ovat samalla rivillä. Chi-square testin voi tehdä helposti R:ssä funktiolla *chisq.test()*. Chi-square testistä saatu p-arvo osoittaa sen, millä todennäköisyydellä saatu tulos on sattumaa. Yleensä todennäköisyyden arvon rajana on 5 % (eli p-arvo on 0.05). Jos tulos on enemmän kuin 5 % ei voi sanoa, että data on varmasti luotettavaa. (Mikhailov 2016, 121-122.)

2.2 Korpuslingvistinen tutkimus

Korpusvetoisen lingvistisen tutkimuksen myötä on tullut mahdolliseksi tutkia kielestä asioita ja siinä esiintyviä kaavoja, joista ei ollut ennen sitä tietoakaan. Tämä soveltuu varsinkin monikielisiin korpuksiin. Voi nähdä miten kielet eroavat tai mitä yhteistä niillä on. Ehkä joskus pystyy jopa näkemään, mikä luonnehtii kieltä yleistasolla. (Johansson 2007, 1.)

Monet tutkimukset ja tutkimustulokset ovat kuitenkin mahdollisia ainoastaan, koska käytössämme on nykyään isoja määriä sähköisessä muodossa olevia tekstejä ja nykyaikaista tietotekniikkaa. Tietotekniikkaa käytetään hyväksi varsinkin tietokonelingvistiikkaan kuuluvalla korpuslingvistiikan alalla, jossa keskitytään lingvististen korpusten (teksti-, puhe- ja sekakorpusten) uusien kokoamis- ja käyttötapojen kehittämiseen. Korpuslingvistiikalla on ainakin kaksi piirrettä, joiden perusteella se voidaan määritellä omaksi opinalakseen. Sen tutkimusmateriaali on sitä itseään varten kehitetty materiaalin muoto eli korpus ja sen käyttöön tarvitaan erityisiä työkaluja. (Zaharov & Bogdanova 2011, 7, 9.)

Korpuslingvistiikan kaksijakoinen luonne (korpusten kokoaminen ja käyttö) näkyy sen kohteen – korpuksen kaksijakoisessa luonteessa. Korpuksethan ovat sekä korpuslingvistiikan materiaalina että sen tuotteena syntyneitä. Korpuslingvistiikan ytimessä ovat juuri korpusten kokoamisen ja käytön teoreettinen pohja ja käytännön toimenpiteet. Korpukset on tarkoitettu laajan käyttäjäpiirin tutkimuksia varten. (Zaharov & Bogdanova 2011, 10.)

Ilman sähköisiä korpuksia lingvistinen tutkimus ei olisi kehittynyt viime vuosikymmeninä niin nopeasti kuin se on. Tutkijoiden saadessa vihdoin käyttöönsä sähköisen korpuksen datan, he pystyivät sen avulla löytämään runsaasti todisteita, jotka tukivat heidän hypoteesejaan kielenkäyttöön liittyen. (Mikhailov & Cooper 2016, 44.)

2.2.1 Korpuslingvistiikkaa edeltäneet lingvistiset tutkimustavat

Korpuslingvistiikkaa voi kuvata alaksi, jolla keskitytään kielen tutkimista varten luotujen proseduurien tai metodien kehittämiseen (McEnery & Hardie 2012, 1). Tekniikat, joita käytetään korpuslingvistiikassa ovat paljon vanhempia kuin tietokoneet: monet niistä ovat 1700- ja 1800-luvuilta, mutta monet ovat myös paljon vanhempia. (Zaharov & Bogdanova 2011, 11). Eli jo kauan ennen korpuslingvistiikan yleistymistä sähköisten korpusten myötä, oli olemassa muita lingvistiikan tutkimustapoja: historiallinen lingvistiikka, kielioppien kirjoittaminen, leksikografia, kieltenopetus ja sosiolingvistiikka. Seuraavaksi kerrotaan näistä tutkimustavoista.

Yksi tärkeimmistä nykyaikaista korpuslingvistiikka edeltäneistä tutkimustavoista oli historiallinen lingvistiikka. Siinä tutkitaan kielen muutoksia historiallisen vertailun metodilla ja ennallistetaan (rekonstruktio) vanhoja kieliä. Ei ole ihmeellistä, että tämä tutkimustapa on vaikuttanut nykyiseen korpuslingvistiikkaan, sillä historiallisen lingvistiikan tutkijat käyttivät ja käyttävät aina aitouden todisteena juuri oikeita tekstejä tai tekstikokoelmia. Monet 1800-luvun tekniikat, joita kehitettiin kantakielten rekonstruktiota ja kieltenvälisten yhteyksien löytämistä varten, ovat käytössä vieläkin. (Zaharov & Bogdanova 2011, 11-12.)

Kielioppitieteilijät havainnollistivat omat väitteensä 1800-luvulla ottamalla esimerkkejä tunnustettujen kirjailijoiden teoksista ja kokoamalla näistä esimerkeistä korpuksia. Nykyään kielioppien tekijät voivat edelleen käyttää hyväkseen korpuksia, mutta nykyaikaiset korpukset sisältävät klassikkokirjallisuuden lisäksi, mitä tahansa muutakin kirjallisuutta. (Zaharov & Bogdanova 2011, 12.)

Leksikologia eli sanasto-oppi on aina liittynyt myös vahvasti korpuksiin:

1700-luvun puolivälissä, kun S. Johnson kirjoitti englannin kielen sanakirjaa (Dictionary of the English language, 1755), hän valitsi kirjoista havainnollistavia lauseita, joita hän kutsui sitaateiksi. Sitaattien avulla hän pystyi näyttämään, miten englanninkieliset kirjailijat käyttivät sanoja. Lukemisen aikana Johnson merkkasi lauseet, joiden konteksti teki sanojen merkitykset erityisen selviksi. Merkatut sanat kopioitiin erillisille papereille, jotka Johnson järjesti sanakirjan kokoamista ja havainnollistamista varten. [oma käännös] (Zaharov & Bogdanova 2011, 13.)

Perinteiset koulujen kieliopit ja oppikirjat sisältävät yleensä keinotekoisesti koottuja tai muokattuja kielenkäytön esimerkkejä. Opiskelijat tulevat ennen pitkää törmäämään aitoihin kielenkäytön tilanteisiin, joissa heille ei välttämättä ole apua näistä keinotekoisista käyttöesimerkeistä. Tästä syystä korpukset ovat hyviä empiirisen ja aidon kielenkäytön tiedon lähteitä. Kieltenopiskelussa korpuksia käytetään muun muassa koneavusteisessa kielenoppimisessa (Computer-Assisted Language Learning, CALL). Se on tietokoneella käytettävä interaktiivinen oppimisympäristö. (Zaharov & Bogdanova 2011, 13-14.)

Sosiolingvistiikka eli kielellinen moninaisuus oli myös yksi lingvistiikan tutkimustavoista, joka edelsi korpuslingvistiikkaa. Se sai alkunsa murrekarttojen ja murreilmaisukokoelmien kokoamisesta 1800-luvun loppupuolella. Sen metodit olivat hyvin samankaltaisia, kuin historiallisen lingvistiikan metodit, mutta suurena erona oli se, että sosiolingvistiikan murrekorpukset koottiin aina joidenkin tiettyjen kriteerien mukaan. Nykyään sähköisiä korpuksia käytetään edelleen kielellisen moninaisuuden (esim. murteiden ja sosiolektien) tutkimisessa. (Zaharov & Bogdanova 2011, 14.)

Huolimatta siitä, että tietokoneet ovat helpottaneet esimerkkien hakua ja luokittelua, ovat monet tekstikorpusten käytön ideat yhä hyvin samanlaisia, kuin ne, joita käyttivät leksikografit ja filologit, joilla ei ollut tietokoneita. Nykyaikainen korpuslingvistiikka käyttää ja kehittää näitä metodeja edelleen (McEnery & Hardie 2012, 1).

2.2.2 Kontrastiivinen tutkimus

Ollakseen hyödyllisiä kontrastiivista tutkimusta varten täytyy korpusten tekstien olla jollain tapaa rinnakkaisia, joko käännössuhteiden takia tai niillä on oltava joku muu rinnastettava ominaisuus (genre, julkaisuvuosi jne.).

Vertailukorpuksen ja käännöskorpuksen yhdistämisen myötä voidaan tehdä Johanssonin (2007) kirjan tutkimusten kaltaisia analyysejä. Hänen kirjassaan käännöksiä käytetään vertailun pohjana ja käännösefektit tarkistetaan vertailukelpoisten alkuperäistekstien perusteella. Vaihtoehtoisesti kontrastiivisessa tutkimuksessa (analyysissä) on perinteisesti käytetty menetelmää: kuvaus,

rinnastaminen ja vertailu, jossa kuvaillaan ja vertaillaan jokaisessa erillisessä kielessä esiintyviä kaavoja pohjautuen vertailukelpoisiin alkuperäiskielisiin teksteihin, ja sitten tutkitaan käännösvastaavuuksia. Metodiikan kannalta merkittävää on se, miten monilla eri tavoilla kaksisuuntaista käännöskorpusta voidaan käyttää. Valittu käyttötapa vaihtelee tutkimuksen päätavoitteen mukaan. (Johansson 2007, 33-34.)

Lingvistisen analyysin metodina, korpuslingvistiikka on sidoksissa kontrastiiviseen tutkimukseen. Kontrastiivinen tutkimus selvittää kieltenvälisiä samankaltaisuuksia ja eroavaisuuksia (Zaharov & Bogdanova 2011, 9). Kontrastiivinen analyysi tai tutkimus on kahden tai useamman kielen järjestelmällistä vertailua. Sen tavoitteena on siis kuvailla niiden eroavaisuuksia ja samankaltaisuuksia. Kontrastiivisella analyysillä voi tutkia joko kielten yleisiä piirteitä tai jollekin kielelle ominaisia piirteitä. Tutkimus voi olla ilman mitään suoraa käyttötarkoitusta tehtyä eli teoreettista, tai se voi olla jotain tiettyä tarkoitusta varten tehtyä eli käytännöllistä. (Johansson 2007, 1.) Kontrastiivisen analyysin suurimpia ongelmia on ekvivalenssi (A on, jos ja vain jos B on) (Johansson 2007, 3).

On vaikea tietää, mitä asioita verrata toisiinsa. Ei ole mahdollista verrata pelkästään kielten muodollisia kategorioita, koska eri kielissä sama ajatus voidaan ilmaista täysin eri keinoin. Korpuskontekstissa käännösparadigma tarkoittaa sitä, miten kohdekielen tietyt kielimuodot vastaavat lähdekielen tiettyjä sanoja tai rakenteita, ja näitä kohdekielen kielimuotoja kutsutaankin *vastaavuuksiksi* (Johansson 2007, 25). Näiden vastaavuuksien tutkimisen avulla voidaan selvittää, mitä asioita voi verrata toisiinsa.

Kieltenväliset vastaavuudet voidaan luokitella sen perusteella, ovatko ne käännöksiä vai alkuperäiskielisiä, esiintyykö vastaavuus liikaa vai liian vähän ja onko se syntaktisesti kongruentti (yhteneväinen) vai divergentti (eriäviä). Kongruentti vastaavuus tarkoittaa sitä, että vastaavuus kuuluu samaan kielelliseen kategoriaan esim. molemmat adverbeja. Divergentti vastaavuus tarkoittaa taas päinvastaista eli vastaavuudet kuuluvat eri kielellisiin kategorioihin. Divergenttien vastaavuuksien perusteella voi nähdä, miten paljon tiettyä tarkoitusta varten käytettyjen kielimuotojen määrä vaihtelee kielten välillä. Näennäisesti vertailtavissa olevien kielimuotojen olemassaolo ei tarkoita sitä, ettei olisi myös divergenttejä vastaavuuksia. Korpus on juuri tästä syystä mahtava työkalu, koska sen avulla voi löytää eroavaisuuksia sieltä, missä ei niitä odottanut olevan. (Johansson 2007, 23-25.)

Kieltenvälisissä vastaavuuksissa on myös mahdollista olla ns. nollavastaavuuksia. Niitä havaitaan yleensä siellä, missä ei ole olemassa luonnollista kieltenvälistä vastaavuutta. Varsinkin

kielimuodoissa, joissa ilmaistaan ihmistenvälisiä ja tekstivälitteisiä merkityksiä. Nollavastaavuudet voidaan jakaa poistoihin tai lisäyksiin. Poisto täytyy kompensoida eli merkitys saatetaan ilmaista jollain muulla kielimuodolla, tai se voidaan jättää pääteltäväksi kontekstista tai jättää kokonaan pois. Lisäyksen voi nähdä kääntäjän käännösratkaisuna, jonka hän perustelee koko tekstin kontekstilla. Lisäyksessä heijastuvat kieltenväliset merkitykset, jotka tavallisesti ilmaistaan luonnollisessa kanssakäymisessä. (Johansson 2007, 26.)

Käännös sopii hyvin kontrastiivisen analyysin tutkimuslähteeksi, koska alkuperäistekstiä ja käännöstä tutkittaessa se näyttää, mitkä kielten osat tai muodot ovat yhteydessä toisiinsa. Useita eri tekstejä ja monien eri kääntäjien tekemiä käännöksiä sisältävät kaksi- tai monikieliset korpukset lisäävät vertailun luotettavuutta ja pätevyyttä. Käännöksiä sisältävän korpuksen käytön voi nähdä jonkinlaisena kääntäjien intuition käyttönä. Tämän intuition perusteellahan korpuksen tekstien lähde- ja kohdekieliset ilmaisut liitetään yhteen. (Johansson 2007, 3-5.)

Aikaisemmat kontrastiiviset tutkimukset olivat tyypillisesti kielten rakenteiden vertailuja. Monikielisen korpuksen avulla voidaan kuitenkin tutkia rakenteiden lisäksi myös sitä, miten niitä rakenteita käytetään oikeassa tekstissä. Vastaavuuksien vertailututkimuksessa selviää, miten usein kielet eroavat siinä, millä tavoin niissä käytetään samanlaisia keinoja tai rakenteita. (Johansson 2007, 6.)

Kontrastiivisen tutkimuksen perusteellinen ongelma onkin vertailtavuus. Johansson (2007) kertookin kirjassaan, että ENPC (English-Norwegian Parallel Corpus) luotiin juuri tämän vertailtavuuden ongelman ratkaisemiseksi. He havainnoivat vastaavuuksia, jotka heijastivat kieltenvälisiä samankaltaisuuksia, jotka kääntäjät havaitsivat. Samalla he tarkistivat, miten hyvin nämä kääntäjien havaitsemat ja valitsemat vastaavuudet vastasivat tavallista kielenkäyttöä. Tällä tavoin he voivat pikkuhiljaa saada selville, mitkä asiat ovat vertailtavissa. (Johansson 2007, 306.)

2.3 Korpusvetoinen käännöstieteellinen tutkimus

Kontrastiivisen analyysin ja käännöksen välinen suhde on kaksisuuntainen. Tiettyjen asioiden käännökset voivat antaa tietoa kontrastiiviselle analyysille. Toisaalta taas kontrastiivinen analyysi voi tarjota selityksiä ongelmiin, joihin törmätään käännöksissä. (Hoey & Houghton 1998, 49.)

Käännöstiede ja kontrastiivinen analyysi ovat erillisiä, mutta toisiinsa liittyviä aloja. Käännöstieteeseen kuuluu monia erilaisia tutkimustapoja. Nämä voivat olla teoreettisia tai deskriptiivisiä käännösilmiöiden tutkimuksia, tai käytännöllisten ongelmien ratkaisua, kuten käännösten arviointi tai kääntäjien koulutus. Eli siis käännöstiede on toisaalta laajempi ja toisaalta

pienempi tutkimusala kontrastiiviseen analyysiin verrattuna. Tämä johtuu siitä, että käännöstiede on rajoitettu tutkimaan käännettyjä tekstejä ja niiden kieltenvälisiä ilmaisutapoja. (Johansson 2007, 4.)

Korpuspohjainen käännöstiede on tähän asti ollut pääasiallisesti kiinnostunut *käännöskorpuksista*. Tässä näillä tarkoitetaan siis Mikhailovin & Cooperin (2016, 220) määritelmää käännöskorpuksille eli yksikielisiä korpuksia, jotka on koottu tietyn kielen käännöksistä ja joita voidaan verrata saman kielen alkuperäisteksteihin. Niistä tutkitaan käännetyn kielen ja tavallisen kielen eroavaisuuksia (käännösefektejä).

Käännös- eli rinnakkaiskorpukset, joista tässä tutkielmassa puhutaan, ovat olleet taas useammin kontrastiivisen tutkimuksen tietolähteinä. Siksi olisikin tärkeää löytää sopiva tasapaino kontrastiivisen tutkimuksen ja käännöstieteen välille, ja lisätä rinnakkaiskorpusten käyttöä käännösprosessien tutkimuksessa. Tähän kuuluisi niiden alueiden ongelmien tutkiminen, joissa kontrastiivinen lingvistiikka ja käännösten lingvistinen teoria kohtaavat.

Käännöstiede on siis erittäin laaja ja monta tiedealaa kattava ala, johon kuuluu kaikki, mikä liittyy kääntämiseen ja tulkkaamiseen. Käännöstieteellisten tutkimusten kohteena voi olla kääntäjien käyttämä kieli, alkuperäistekstin ja käännöksen suhde, itse käännösprosessi tai käännöksen automatisaatio (Kenny 1998, 51). Tästä voi siis päätellä, että rinnakkaiskorpukset soveltuvat usein todella hyvin tällaisten tutkimusten tietolähteiksi.

2.3.1 Konekäännös

Kun mainitaan käännökset ja tietokoneet samassa yhteydessä, tulee usein mieleen konekäännös ja sen ongelmat. Termi *konekäännös* tarkoittaa tietokonejärjestelmää, joka kykenee tuottamaan käännöksiä ihmisen avustamana tai avustamatta. Siihen eivät kuitenkaan kuulu tietokoneisiin pohjautuvat käännöstyökalut, jotka tukevat kääntäjää antaen hänelle mahdollisuuden päästä käsiksi netissä oleviin sanakirjoihin, termitietopankkeihin yms. Koneen avustaman ihmisen tekemän käännöksen (machine-aided human translation) ja ihmisen avustaman konekäännöksen (human-aided machine translation) rajat ovat usein epäselvät, ja niistä molemmat voidaan luokitella osaksi tietokoneavusteista kääntämistä (computer-aided translation). Konekäännös eroaa näistä kuitenkin siinä, että konekäännöksessä keskitytään koko käännösprosessin automatisaatioon. (Hutchins 1994, 2322.)

Jo 1930-luvulla oli monia tutkijoita, jotka uskoivat automaattisten koneiden auttavan kielten kääntämisessä. Vuonna 1933 ranskalainen insinööri George Artsuni patentoi käännöskoneen, jota hän kutsui mekaanisiksi aivoiksi. (Heinsz-Dostert & Macdonald, Zarechnak 1979, 7.)

Informaatio- ja viestintäteknologiat ovat kehittyneet nopeasti viime vuosikymmeninä. Konekäännös on yksi tärkeimmistä ja vaikeimmin toteutettavista aloista. Tämä johtuu siitä, että onnistuakseen järkevällä tavalla, konekäännös tarvitsee erittäin monimutkaisia prosesseja. (Alsohybe & Dahan & Ba-Alwi 2017, 2, 4).

Netin kautta toimivilla käännösohjelmilla voi kääntää isojakin tekstejä. Tuloksena saadaan kuitenkin usein huonoja käännöksiä, joista kuvastavat konekäännösten ongelmia ja monimutkaisuutta. Viime vuosikymmenten aikana konekäännösohjelmissa on käytetty yhä enemmän ja enemmän hyödyksi korpuspohjaisia metodeja. Tekstidatan ensimmäinen käyttötarkoitus oli konekäännösjärjestelmien tehokkuuden testaaminen. Myöhemmin tutkijat alkoivat käyttää rinnakkaistekstejä ihmisen ja koneen tekemien käännösten vertailuun. Lopulta alkoi ilmestyä tilastolliseen tietoon perustuvia automaattisia kääntäjiä. Ne perustuvat rinnakkaisteksteihin ja niistä saatuihin esimerkkeihin, joten niitä kutsutaan nimellä *Example-based Machine Translation* (esimerkkeihin perustuva konekäännös). Yksi esimerkki tällaisesta ohjelmasta on Google Translate eli Google Kääntäjä, jota meistä jokainen on varmasti joskus kokeillut käyttää. Tällaiset ohjelmat eivät kuitenkaan ole ammattikääntäjien käyttöä varten, vaan ne auttavat tavallisten ihmisten päivittäisissä käännöstarpeissa. (Mikhailov & Cooper 2016, 184-185.)

Konekäännösohjelmat eivät siis ole ammattikääntäjien käyttöä varten, koska ne eivät pysty tuottamaan korkeatasoista käännöstä ilman ihmisen apua. Konekäännöksiä varten olevat tekstit täytyy joko kirjoittaa yksinkertaistetusti tai muokata etukäteen muulla tavoin. Myös konekäännöksestä saadut tulokset täytyy muokata jälkikäteen ihmisen toimesta. Konekääntämisen käyttötarkoitus onkin usein käyttää niitä yhdessä käännösmuistiohjelmien kanssa avustamaan ihmiskääntäjää, eikä niinkään kääntämään suoraan itsestään. (Mikhailov & Cooper 2016, 185.)

2.3.2 Kääntäjät & käännösmuistiohjelmat

Rinnakkaistekstit eivät itsessään ole niin harvinaisia kuin saattaisi olettaa. Vanhimmat niistä ovat tuhansia vuosia vanhoja. Rinnakkaistekstejä ovat esim. kaiverretut kirjoitukset kahdella tai kolmella kielellä (esim. Rosettan kivi), kaksikieliset sopimukset, klassiset tekstit ja niiden modernit käännökset sekä monikieliset käyttöohjeet ja nettisivut. (Mikhailov & Cooper 2016, 197.)

Kääntäjät tuottavat tekstejä monia eri tarkoituksia varten. Useimmiten nämä tekstit kuuluvat samaan genreen tai niissä käsitellään samoja aiheita. Tämänkaltaiset tekstit sisältävät samoja termejä, fraaseja ja samanlaisia sanamuotoja. Jotkut niistä ovat samanlaisia myös kokonaisrakenteeltaan ja hyvin kaavamaisia tai niissä on jopa sama pohja. Tällaisia ovat esimerkiksi CV, syntymätodistus tai

sopimukset. Samaa pohjaa käyttävät saattavat olla lähes identtisiä ja niissä on muutettu ainoastaan henkilöiden nimet ja päivämäärät. (Mikhailov & Cooper 2016, 41.)

Tämmöisissä tapauksissa kääntäjät haluavat luonnollisesti aikaa säästääkseen käyttää omia aiempia käännöksiään tai muiden kääntäjien tekemiä käännöksiä. Tämä takaa myös käännösten johdonmukaisen tyylin ja terminologian. Käännösmuistiohjelmat kehitettiin tähän tarkoitukseen. Kääntäjä luo tekstihakemiston sijaan erillisen tekstiarkiston, joka tallennetaan tietokantana, johon pääsee käsiksi käännösmuistiohjelmien avulla. (Mikhailov & Cooper 2016, 13.)

Käännösmuisti on siis arkisto, jossa on aiemmin käännettyjä tekstejä jaettuna segmentteihin. Tavallisesti tämmöinen segmentti on virke, otsikko tai listan osa. Lähdekielen tekstisegmentit kohdistetaan kohdekielen tekstisegmentteihin, sillä tavoin, että niitä voidaan kierrättää käännösmuistityökalun avulla. Käännösmuistityökalu ohjaa käännösprosessia, tarjoten käyttäjälle käyttöliittymän, jossa voi nähdä lähde- ja kohdekielen tekstisegmentit. Työkalu luo lisäksi käännösmuistia jatkuvasti käännettäessä, jolloin se tallentaa lähde- ja kohdekielen tekstisegmentit käännösyksiköiksi. (Moorkens 2013, 31.)

Käännösmuisti luodaan tavallaan lennosta samalla kuin käännetään. Kääntäjien on kuitenkin mahdollista lisätä heidän aiempia käännöksiään käännösmuistiohjelmiin. Käännösmuistiohjelmat eivät ole sama asia kuin konekääntäminen. Ne eivät siis käännä, vaan tuottavat otteita valmiiksi olemassa olevista käännöksistä. Käännösmuistit ovat myös erilaisia, verrattaessa tekstiarkistoihin. Niiden tarkoituksena ei ole antaa tietoa eikä niiden kautta voi käsitellä kokonaista tekstiä, koska ne eivät sisällä kokonaisia tekstejä, vaan kohdistettujen tekstisegmenttien kokoelmia, joita käyttäjä voi halutessaan muokata tai poistaa kokonaan. (Mikhailov & Cooper 2016, 14-15.)

Voiko käännösmuistia sitten kutsua korpukseksi? Korpusten tavoin ne ovat tekstitietoa sisältäviä tekstikokoelmia ja sisältävät rinnakkaiskorpusten tavoin suuria määriä kohdistettuja tekstisegmenttejä. Ne ovat myös päivitettävien korpusten tavoin avoimia. On kuitenkin vaikea sanoa, että ovatko ne korpuksia. Avoimuutensa takia niistä on hyvin epätodennäköistä saada hyötyä lingvististä tutkimusta varten. Toisaalta niiden kohdistustyökalulla voidaan luoda rinnakkaiskorpus. (Mikhailov & Cooper 2016, 14-15.)

Käännösmuistit muistuttavat siis rinnakkaiskorpuksia, mutta eroavat kuitenkin niistä esim. siten, että ne syntyvät luonnollisesti käännösprosessin tuotteena. Ne eivät myöskään sisällä toistuvia tekstisegmenttejä, vaan tämmöisten vastaan tullessa, käännösmuisti ehdottaa kääntäjälle aiemmin käytettyjä käännöksiä (Moorkens 2013, 31).

3. KONDITIONAALI SUOMESSA JA VENÄJÄSSÄ. MITÄ TIEDETÄÄN?

Modaalisuus on semanttinen alue, jossa on kyse asiaintilan todenmukaisuutta ja toteutumismahdollisuuksia koskevista arvioista. Modaalisilla kielenaineksilla puhuja ilmaisee, onko asiaintila hänen mielestään tai yleisesti ottaen varma, välttämätön, todennäköinen, mahdollinen, epävarma tai mahdoton, pakollinen tai luvallinen, toivottava tai epätoivottava, ulkoisista tai sisäisistä edellytyksistä riippuvainen. (VISK § 1551.)

3.1 Konditionaali suomessa

Verbeillä on moduksia eli tapaluokkia, jotka ovat verbien taivutuksen kategorioita. Moduksilla ilmaistaan episteemistä (potentiaali ja konditionaali) ja deonttista (imperatiivi) modaalisuutta. Näistä kolmesta konditionaalia käytetään eniten muuhun tarkoitukseen kuin pelkästään modaalisuuden ilmaisuun. Konditionaalin merkityksessä on siis mahdollisuus, mutta erilainen kuin muissa modaalisuuden ilmaisukeinoissa: tarkastelussa oleva asiantila ei ole todenmukaisena vaan se riippuu puhujan tahdosta tai mielikuvituksesta eli se on vaihtoehtoinen. Konditionaali kuvastaa siis sitä, että milloinkin kyseessä oleva asiantila on yhtenä vaihtoehdoista, eikä tosi. (VISK §1590.)

Suomen kielessä konditionaali muodostetaan verbistä ja sen tunnuksesta *-isi-*. Sillä on kaksi aikamuotoa (tempusta): preesens (*selviäisi*) ja perfekti (*olisi selvinnyt*). (Finnlectura.fi 2.5.2.1.4.)

Konditionaali ilmaisee tekemisen ehdollisena tai epävarmana. Sillä voi myös ilmaista halua, toivomusta, suostuttelua, epäilyä tai kohteliasta pyyntöä. Potentiaalista poiketen, konditionaali on yleisesti läsnä sivulauseissa. Sen käyttö on runsasta varsinkin ehtoa ilmaisevissa jos-lauseissa. (esim. *Jos olisin kuningas, asuisin linnassa*.) Jos-lauseella esitetään ehto, jonka voimassa ollessa se propositio (ehdotus), joka on nyt todenvastainen, pitää paikkansa. Kun referoidaan että-lauseita konditionaalin avulla, ilmaistaan sitä, että puhuja ei esitä ehdotustaan totena, vaan jättää asian todenperäisyyden avoimeksi. (esim. *Väitätkö, että valehtelisin tästä sinulle?*). Konditionaalilla voi suomen kielessä referoida myös imperatiivia, jolloin se vastaa muodoltaan alkuperäisen rakenteen käskymuotoa (esim. *Marko pyysi, että hakisin vettä kaivosta*). (Finnlectura.fi 2.5.2.1.4.)

Referoitaessa jonkun henkilön tahtoa ja aikeita on käytössä 3. persoonan konditionaalimuoto (esim. *Kilpailu laskisi hintaa*.) Referoivassa sivulauseessa voidaan konditionaalilla osoittaa vastaavasti epäilevää suhtautumista alkuperäiseen lausumaan. (esim. *Lehtijutut siitä, että kilpailu laskisi hintaa, ovat täysin perusteettomia*). (VISK § 1592.)

Konditionaali on suunnittelun, kuvitellun ja ennustamisen modus. Suunnitelmissa oleva tai intensionaalinen asiaintila on mahdollinen ajateltavissa oleva vaihtoehto. (esim. *Näistä saappaista saisi hyvät kalossit*.) Konditionaalia käytetään myös ennusteissa, jolloin mahdollinen asiaintila esitetään suhteutettuna johonkin kuviteltuun tilanteeseen. Ennustetulkintainen konditionaalilause siis esimerkiksi jatkaa jotain oletusta ja muodostaa sen perusteella seurauksia tai muita asiaintiloja. (esim. *Uskon, että sen toteuttaminen tulisi hyvin kalliiksi*.) (VISK § 1592.)

Konditionaalia käytetään suomen kielessä myös vetoamisen ja ehdottamisen keinona, jolloin sitä kutsutaan tahtotulkintaiseksi konditionaaliksi. Se esiintyy vetoomuksissa, tiedusteluissa, ehdotuksissa ja kutsuissa. Tahtotulkintainen konditionaalilause on olennainen jonkin tiedossa olevan suunnitelman tai aikomuksen kannalta (esim. *Kävisit nyt siellä jo*.) Konditionaalin preesensin tai perfektin sisältävällä lauseella voidaan ottaa huomion kohteeksi joku kontekstissa oleva asia. Verbin konditionaalimuodon lisäksi predikatiivilauseen tai e-lauseen preesens nähdään tarjouksena tai muuna ehdotuksena. (esim. *Siellä olisi nyt ruokaa*.) Perfektillä taas puolestaan esitetään menneille tapahtumille vaihtoehtoinen kulku. (esim. *Tuossa olisi ollut kahvila*. (mentiin jonnekin kauemmas kahville, kun ei huomattu sitä)). Konditionaalia voidaan käyttää myös asioimistilanteissa, jolloin konditionaalimuodon sisältävä puheenvuoro eroaa jollain tavalla normista ja edellyttää jotain, esim. vaihtoehtojen käsittelyä: *Uutta puhelinta olisin katsellut.* (Edellyttää myyjältä vaihtoehtojen käsittelemistä ja tarjoamista). (VISK § 1593.)

Konditionaalilla on lisäksi todenvastaisen ja toteutumattoman modaliteetin tulkinta. Tällä tavoin tulkitaan preesensin ja perfektin konditionaalimuodot esimerkiksi retorisissa kysymyksissä sekä kieltolauseille alisteisissa kysymyksissä esim. *En muista, milloin olisin nukkunut näin sikeästi*. Lisäksi konditionaalin perfekti ilmaisee menneisyyteen sijoittuvan asiantilan toteutumatta jääneen vaihtoehdon esim. *Olisit heittänyt pidemmälle*, tai vielä toteutumattoman esim. suunnitellun vaihtoehdon *Olisi pitänyt korjailla pyörää vähän tässä*. (VISK § 1593.)

Suomen kielessä konditionaalia käytetään myös ehtolauseissa. Ehtolauseita käytetään, kun puhutaan asioista, jotka riippuvat toisistaan ja ovat tavalla tai toisella yhteenkuuluvia. Lauseissa, jotka ilmaisevat ehtoa ei ole välttämättä *jos*-partikkelia: *Kuulisit tuon kappaleen niin tietäisit, miksi pidän siitä.* (Jos kuulisit). Ehtolauseen verbin ollessa indikatiivissa, sitä seuraavan lauseen konditionaali voidaan tulkita suositukseksi tai ehdotukseksi, eikä sillä ilmaista silloin samalla tavalla riippuvuutta edellä esitetystä ehdosta kuin jos lauseilla on sama modus. *Jos lihakset menevät vielä enemmän jumiin, sinun pitäisi käydä hierojalla.* Jos-lauseen sijaan voidaan ehto ilmaista substantiivilausekkeella: *Kolme kertaa enemmän puuroa, niin se olisi riittänyt kaikille.* Ehtolauseet

voivat olla myös kontrafaktuaalisia, jolloin tiedetään, että ehto on todenvastainen: *Jos tämä olisi helppoa, olisivat kaikki mestareita.* (VISK § 1595.)

Konditionaalimodus esiintyy lisäksi finaalisissa, konsessiivisissa ja konditionaalisissa adverbiaalilauseissa, joilla ilmaistaan todenvastaista tai toteutumatonta asiantilaa. *Jos – niin* -yhdysrakenteessa voi konditionaalimuoto olla sekä ehtoa ilmaisevassa *jos*-lauseessa että seurausta ilmaisevassa *niin*-lauseessa. Muista yhdyslauseista ehtorakenteet eroavat sillä tavoin, että lauseet eivät ota kantaa siihen, onko kyseessä oleva tapahtuma toteutunut tai toteutumassa. Ne osoittavat vain, miten yhdessä lauseessa olevan ehdon (propositio) todenmukaisuus riippuu toisen lauseen proposition todenmukaisuudesta. (VIKS § 1334.)

3.2 Konditionaali venäjässä

Venäjän kielessä konditionaali ilmaistaan partikkelin *by (b)* avulla. Konditionaaliin kuuluvat partikkelin *by (b)* erilaiset yhdistelmät (konjunktiot) kuten *čtob(y)* (vastaa suomen kielessä: että, jotta) ja vanhentuneet *daby* sekä *kaby*. Konditionaalin muodostava partikkeli *by* voi siis esiintyä täydessä tai lyhennetyssä muodossa *b*, tai yhdistelmänä *čtob(y)*. Täydellä muodolla *by* ei ole morfologisia rajoitteita yhdistelylle, mutta lyhennetyllä muodolla *b* on muodollisia rajoitteita käyttökontekstin suhteen: *b* ei voi käyttää sellaisen sanan jälkeen, joka päättyy konsonanttiin. (Dobrušina 2014, 68.)

Venäjän kielen konditionaali ei erota aikamääreitä (Russkaâ grammatika 1980, 625). Tilanne, jota konditionaalimuoto kuvastaa, voi siis kuvastaa samanaikaisesti preesensiä, perfektiä ja futuuria.

Partikkeli *by (b)* voi esiintyä yhdessä verbien perfektimuotojen kanssa, verbin infinitiivimuotojen kanssa, predikatiivien kanssa, elliptisissä rakenteissa, objektiivisessa sijassa olevien substantiivien kanssa, muiden verbiin liittymättömien yksiköiden yhteydessä, partisiippien kanssa, gerundien kanssa ja imperatiivien kanssa. (Dobrušina 2014, 68.)

Näistä huomattavasti yleisimpiä ovat verbin perfektimuodon ja partikkelin *by (b)* yhdistelmät (Dobrušina 2014, 68). Tätä kutsutaan usein ***-l-*** muodoksi, koska sillä on sama ulkomuoto, kuin aidolla verbin perfektillä, esim. *igral by* (Russkaâ grammatika 1980, 625). Näin voidaan erottaa tämä konditionaalin kanssa esiintyvä perfektimuoto (***-l-*** muoto) oikeasta verbin perfektimuodosta, koska konditionaali ei erota aikamääreitä, joten tämä verbin perfektimuoto menettää mennyttä aikaa kuvaavan merkityksensä. Kuitenkin konditionaalimuodon kuvauksessa on otettu tavaksi kutsua tätä ***-l-*** muotoa perfektimuodoksi. (Dobrušina 2014, 68.)

Näistä syistä tässäkin tutkielmassa, kun puhutaan **venäjän konditionaalin perfektimuodosta**, tarkoitetaan juuri tätä ***-l-* muotoa**.

Todellisessa konditionaalimuodon kuvauksessa kaikki säännöllisesti partikkelin *by (b)* kanssa esiintyvät muodot luetaan konditionaalimuodoiksi, koska funktionaalisesti muodot, joissa käytetään verbin perfektimuotoa ja muut rakenteet, joiden kanssa *by (b)* esiintyy, ovat samankaltaisia (konditionaalia esittäviä). Jos *by (b)* esiintyy sellaisten rakenteiden kanssa, jotka eivät venäjässä kuulu konditionaalin alaisuuteen, niin ne eivät ole konditionaalimuotoja, vaikka ne olisivat *by (b)* partikkelin kanssa kirjoitettuna. Esimerkiksi *vrode by uhodit* eli *on olevinaan lähdössä*. (Dobrušina 2014, 69-70.)

Venäjän kielessä konditionaali kuvastaa tilanteet, joita ei ole olemassa oikeassa (reaalisessa) maailmassa. Se voi olla merkityksiltään: kontrafaktuaalinen (esim. *На твоем месте я бы этого не делал*: Sinun tilallasi en tekisi noin) ja toivottavuutta ilmaiseva (esim. *Только бы он не заметил*: Kunpa hän ei vain huomaisi). Kuvaannollisessa käytössä konditionaalilla on myös muita käytännöllisiä funktioita: kertoa puhujan tarkoituksista lievennetysti (esim. *Я бы попросил вас не говорить об этом*: Pyytäisin teitä olemaan puhumatta siitä) ja kun yritetään vähentää jonkin väitteen jyrkkyyttä (esim. *Я назвал бы это воровством*: Minä sanoisin sitä varkaudeksi). Konditionaalia käytetään yleisesti erialisissa sivulauseissa, esimerkiksi: ehtolauseissa, alisteisissa ehtolauseissa (esim. *Сколько бы он не говорил, я все равно его не понимал*: Olisi hän puhunut kuinka paljon tahansa, en silti tajunnut häntä), tavoitteellisissa lauseissa (esim. *Иди тихо, чтобы не кто не услышал*: Mene hiljaa, jottei kukaan kuulisi), lisäyksen tekevissä lauseissa (esim. *Марко хочет, чтобы ты не кому не рассказывал*: Marko haluaa, että et kertoisi kenellekään) ja relatiivilauseissa. (Dobrušina 2014, 66.)

Konditionaalimuodolla on monta merkitystä, jotka ovat tyypillisiä epärealistisille muodoille eli se kuvaa tilanteita, joita ei ole reaalimaailmassa. Tätä kutsutaan kontrafaktuaaliseksi merkitykseksi, joka on venäjän kielen konditionaalimuodon varsinainen merkitys. Kontrafaktuaalisiksi tilanteiksi kutsutaan tilanteita, jotka eivät ole olleet olemassa reaalimaailmassa eivätkä tule koskaan olemaan, mutta joita puhuja tarkastelee kuin asioita, jotka kuuluisivat johonkin toiseen maailmaan. Tätä merkitystä ilmenee sekä kerronnallisessa tekstissä että dialogissa. Kaksi muuta merkitystä ovat toivottavuus (toivottavaa asiaa ilmaiseva merkitys) ja lievennys (sanotun asian jyrkkyyttä vähentävä merkitys). (Dobrušina 2014, 71.)

Toivottavuutta ilmaisun tilanteita ovat sellaiset, jotka ovat olemassa reaalimaailmassa ja joiden todenmukaisuuden puhuja kokee positiiviseksi (toivottavaksi). Verbin perfektistä muodostuva

konditionaalimuoto ei sinällään sisällä mitään arvostelukomponenttia. Puhujan positiivinen arvostelu kyseistä asiaa kohtaan tulee ilmi kontekstista, partikkeleista ja adverbeista. Toivottavuutta ilmaiseva konditionaalimuoto ilmenee perfektissä olevan konditionaalimuodon kanssa huomattavasti harvemmin, kuin konditionaalimuotojen kanssa, jotka muodostuvat infinitiivistä, predikatiivista, substantiivista tai muista yksiköistä. Lievennystä ilmaiseva merkitys ilmenee ainoastaan dialogissa ja sen tarkoituksena on lieventää sanojan tai sanotun asian jyrkkyyttä. (Dobrušina 2014, 73.)

3.3 Konditionaalien vertailu

Konditionaalin merkityksessä on eroavaisuuksia suomen ja venäjän välillä. Suomessa konditionaali ilmaisee tarkasteltavan asiaintilan vaihtoehtoisena eli puhujan tahdosta tai mielikuvituksesta riippuvana, yhtenä vaihtoehtona. Venäjän kielessä konditionaali taas ilmaisee asiantilaa, jota ei ole olemassa reaalimaailmassa. Venäjässä asiantila ei siis voi olla eikä tulla koskaan olemaan oikeassa maailmassa ja suomessa se on yksi mahdollisista vaihtoehdoista. Toki suomen kielessä konditionaalilla on myös todenvastaisen tai toteutumattoman modaliteetin tulkinta, mutta se on vain yksi pienempi osa konditionaalin merkitystä eikä sen päätarkoitus, kuten venäjässä.

Suomen kielessä konditionaali voi ilmaista ehdollisuutta, epävarmuutta, halua, toivomusta, suostuttelua, epäilyä tai kohteliasta pyyntöä. Venäjässä konditionaalilla ilmaistaan jonkin toivottavuutta, lievennetään sanotun tai sanojan jyrkkyyttä sekä ilmaistaan tilanteita, joita ei ole olemassa oikeassa maailmassa. Venäjässä konditionaalia käytetään lisäksi ilmaisemaan ehtoa, tavoitetta tai lisäystä.

Suomen kielessä konditionaalin tunnus on *-isi-* ja sillä on kaksi aikamuotoa: preesens ja perfekti. Venäjässä konditionaalin tunnus on partikkeli *by (b)* ja se ei erota aikamääreitä vaan tilanne, jota konditionaali kuvastaa voi olla samanaikaisesti perfekti, preesens ja futuuri eli sillä ei ole oikeastaan aikamuotoa.

Venäjän kielessä konditionaali on rakenteeltaan huomattavasti vapaampi. Esimerkiksi partikkelilla *by (b)* ei täydessä muodossa eli *by* ole morfologisia rajoitteita yhdistelylle. Suomen kielessä konditionaalilla on tunnus *-isi-*, joka voi esiintyä ainoastaan preesens tai perfektimuotoisen verbin kanssa.

3.4 Hypoteesini teorian perusteella

Suomen ja venäjän kielen konditionaalimuodoissa on siis eroavaisuuksia. Joissakin tapauksissa, kuten sellaisissa, missä suomen kielen konditionaalilla ilmaistaan todenvastaista tai toteutumatonta

modaliteettia, voidaan mielestäni kääntää suomenkielinen konditionaalimuoto venäjän konditionaalimuodolla. Kieltenväliset konditionaalimuodot korreloivat myös muissa tilanteissa, kuten sellaisissa, joissa ilmaistaan ehtoa, halua, toivetta. Molemmissa kielissä konditionaalia käytetään myös lieventämään pyyntöä. Konditionaalin päämerkitys on kielissä kuitenkin hyvin erilainen: venäjän kielessä konditionaalin päätarkoitus on todenvastaisen tai toteutumattoman modaliteetin ilmaisu ja suomen kielessä se on vain yksi pienempi osa konditionaalin merkitystä.

Teoriaosuuden pohjalta rakennettu hypoteesini on seuraava: **Konditionaalin päämerkitys on kielissä erilainen, joten oletan, että suomen kielen konditionaalimuoto on käännetty suurimmaksi osin jollain muulla tavalla, kuin venäjän kielen konditionaalimuodolla. En osaa vielä sanoa, mikä on yleisin tapa kääntää suomen kielen konditionaali venäjään. Venäjän kielen konditionaali rakentuu useimmiten verbin perfektimuodolta näyttävästä *-l-* muodosta ja suomenkielinen konditionaali voi myös olla aikamuodoltaan perfekti, joten oletan, että perfektiaikamuotoinen konditionaali kääntyy venäjään useammin konditionaalilla kuin preesensaikamuotoinen.**

4. SUOMEN KONDITIONAALIN VASTINEITA KORPUSAINEISTOSSA

4.1 ParFin korpus & aineiston haku

Tutkimusaineistona tutkielmassa on käytetty aineiston lähteenä Tampereen yliopiston käännöstieteen tutkijoiden laatimaa kaunokirjallisuuden suomi-venäjä paralleelikorpusta (rinnakkaiskorpusta) *ParFin*, jonka kokoamisen tohtoriopiskelija Juho Härme aloitti vuonna 2010. Korpuksesta löytyvää aineistoa on käytetty sellaisena kuin se on ollut keväällä 2018.

Tampereen yliopiston rinnakkaiskorpusten *ParRus* (venäjä-suomi) ja *ParFin* (suomi-venäjä) kokoaminen on aloitettu eri aikoina. ParFin korpuksen kokoaminen aloitettiin vuonna 2010, siis huomattavasti myöhemmin kuin ParRus korpuksen, jonka kokoamisen Tampereen yliopiston professori Mikhail Mikhailov aloitti jo vuonna 1999. Suunnitelmissa on, että molemmat korpukset yhdistetään jossain vaiheessa yhdeksi kaksisuuntaiseksi venäjä-suomi-venäjä korpukseksi. Näiden korpusten yhdistämisessä on kuitenkin sama ongelma, kuin monilla muillakin esimerkiksi ENPC (English-Norwegian Parallel Corpus) (kts. Johansson 2007). On paljon enemmän käännöksiä venäjästä suomeen kuin suomesta venäjään. (Mikhailov & Cooper 2016, 33, 208-209.)

Tampereen yliopistolla koottuja korpuksia varten on kehitetty, jatkuvasti työstettävä ohjelmistopaketti nimeltä Texthammer. Sitä kehittävät Mikhail Mikhailov ja Juho Härme. Sen avulla tehdään hakuja eri korpuksiin, jotka ovat tallennettuina mustikka.uta.fi palvelimelle. Texthammer ohjelmistoa kehitetään, jotta sen kautta olisi mahdollista päästä käsiksi sekä yksikielisiin korpuksiin että rinnakkaiskorpuksiin nettiselaimen avulla. Korpukset on tallennettu palvelimelle Postgrsql tietokantoihin. Texthammer ohjelmisto käyttää PHP skriptejä, jotka tekevät SQL (Structured Query Language on IBM:n kehittämä standardisoitu kyselykieli) kyselyitä tietokantoihin, joissa korpuksen data sijaitsee, ja näyttää hakutulokset nettiselaimessa. Sovelluksen pääfunktio on tehdä erilaisia hakuja. Sitä ei ole suunniteltu hakutulosten käsittelyyn. Tästä syystä on parempi ladata hakutulokset texthammerista, johonkin sovellukseen, jolla hakutulokset voi käsitellä. Käsittelyn voi tehdä erilaisilla taulukko- ja tietokantaohjelmilla. Esimerkiksi tämän tutkielman aineisto käsiteltiin Microsoft Excelillä ja R:llä. (Texthammer, ver. 1.5. User manual.)

Korpuksesta haettiin osakorpus, johon kuuluivat kaunokirjalliset tekstit, jotka on julkaistu vuoden 1980 jälkeen ja niiden käännökset. Näistä haettiin 1000 esimerkkiä, joissa suomenkielisessä esimerkissä verbi on konditionaalimuodossa. Haku tapahtui mustikka.uta.fi palvelimen Texthammer ohjelmiston paralleelikonkordanssihaulla. Koska ParFin korpus on morfologisesti annotoitu, on siitä mahdollista hakea kieliopillisia muotoja. Haussa haettiin siis konditionaalimuotoja kirjoittamalla

rinnakkaiskonkordanssihaun *Grammar form* hakukenttään: %Mood=Cnd%. Tällä haulla tulokseksi saadaan suomenkielisiä esimerkkejä, joissa on käytetty konditionaalimuotoa ja niiden rinnakkaisia käännösvastineita venäjän kielellä. Esimerkit konditionaalin käytöstä haettiin lisäksi satunnaisessa järjestyksessä, jotta ne eivät olisi kaikki yhdestä tai kahdesta teoksesta otettuja. Tällä tavoin kaikki esimerkit eivät ole myöskään vain yhden tai kahden kääntäjän kääntämiä.

4.2 Metodien avaus

4.2.1 Monimuuttujatilastot

Aineistoni perusteella päätin, että kyseisessä tutkimuksessa paras tutkimustapa on käyttää monimuuttujatilastometodeja. Muuttuja on abstrakti konsepti (ikä, paino…), joka on suhteessa johonkin tietyn objektin (henkilö, sana…) tiettyyn piirteeseen, joka on tutkijan mielestä tutkimukselle relevantti. Arvo on taas konkreettinen esimerkki tästä abstraktista konseptista, jossain tietyssä tilanteessa. (De Sutter 2018, 33.)

Muuttujia on erilaisia. On olemassa esimerkiksi jatkuvia muuttujia, jotka voivat esiintyä, missä tahansa numeerisen skaalan datapisteessä esim. pituus. Epäjatkuvat muuttujat voivat esiintyä taas ainoastaan tietyssä skaalan datapisteessä esim. henkilön paino. (Martin & McFerran 2017.) Oman tutkielmani muuttujat ovat epäjatkuvia muuttujia. On lisäksi olemassa tutkimuksen kannalta tärkeitä eri muuttujia: päämuuttuja eli tutkimuksen keskeinen muuttuja ja sitä voi kutsua nimellä riippuva muuttuja tai **vastemuuttuja**. Ne muuttujat, joiden tutkija epäilee vaikuttavan jollain tapaa vastemuuttujaan, ovat itsenäisiä tai **selittäviä muuttujia**. Esimerkiksi omassa tutkimuksessani: Yhtenä selittävänä muuttujana sille, että suomen kielen konditionaalimuoto on käännetty konditionaalilla venäjään, voi olla itse kääntäjä. Tällöin vastemuuttujana olisi käännöksen tapaluokka ja selittävänä muuttujana kääntäjä.

Näiden lisäksi on olemassa sekoittavia muuttujia (confounding variable) eli muuttuja, joka vaikuttaa sekä selittävään muuttujaan että vastemuuttujaan ja voi saada tutkijan analysoimaan saadun tuloksen väärin (Shuttleworth & Wilson 2008). Esimerkki sekoittavasta muuttujasta: muuttujilla X ja Y on sekoittava muuttuja Z silloin, kun Z voi olla syynä sekä X:lle että Y:lle. Eli jos tutkitaan esim. selittävän muuttujan *aktiivisuustaso* vaikutusta vastemuuttujaan *painon lisääntyminen*, niin saatetaan huomata merkittävä korrelaatio näiden välillä. Näillä muuttujilla on kuitenkin sekoittava muuttuja *ikä*, joka vaikuttaa sekä aktiivisuustasoon että painon lisääntymiseen.

Sekoittavien muuttujien etsimisen voi tehdä tarkastelemalla visuaalisesti muuttujien suhdetta, mutta tämä voi olla hankalaa ja aikaa vievää. Tässä apuun tulevat monimuuttujatilastomenetelmät, joista esim. logistinen regressioanalyysi voi auttaa tutkijaa. (De Sutter 2018, 22.)

Tilastoja tarvitaan siihen, että voidaan arvioida, millä varmuudella voidaan yleistää tuloksia, jotka on saatu jostain näytteestä, kuvastamaan koko otantakantaa. Monimuuttujatilastollisetmenetelmien avulla voi arvioida kahden tai useamman kilpailevan itsenäisen muuttujan suhteellista vaikutusta vastemuuttujaan. Ne voivat auttaa löytämään tärkeät ja turhat tekijät, jotka vaikuttavat saatuun tulokseen. Empiirisen tutkimuksen parissa tutkija ottaa yleensä näytteen isommasta materiaalista (otantakannasta). Joten yksi tärkeimmistä toimenpiteistä, joka kuuluu lähes kaikkeen empiiriseen työhön, on induktiivinen päättely eli induktio. Se on päättelymuoto, jossa yleistetyt väittämät ja päätelmät, joita tehdään koko otantakantaan, perustuvat rajoitettuun määrään havaintoja eli näytteeseen. Tilastot auttavat tässä toimenpiteessä. Kuvailevat eli deskriptiiviset tilastot (kuvaileva analyysi) auttavat summaamaan näytteessä tapahtuvat variaatiot ja erottamaan päätaipumukset. Tilastollinen päättely taas auttaa selvittämään, kuinka luotettavia rajoitetuista havainnoista tehdyt päätelmät ovat. (De Sutter 2018, 25-32.)

Tilastollinen päättely auttaa siis selvittämään, kuinka luotettavia rajoitetuista havainnoista tehdyt yleistykset ovat. Toisin sanottuna, kuinka varma voi olla siitä, että päätelmät, jotka on tehty osittaisen informaation (näytteen) perusteella, edustavat koko otantakantaa. Tilastollinen päättely tunnetaankin laskelmoidun riskin tieteenä. Tärkeitä konsepteja tilastollisen päättelyn suhteen ovat **nollahypoteesi** ($H_0 : x = 0$) ja **p-arvo**. Kaikissa tilastollisissa testeissä on olettamuksia kunkin muuttujan teoreettisesta jakaumasta. Kun kyseessä on tietty näyte, on nollahypoteesi seuraava: kyseinen muuttuja seuraa teoreettista jakaumaa eli näytteessä ei tapahdu mitään erikoista. Vastaavasti on olemassa myös vaihtoehtoinen hypoteesi ($H_a : x \neq 0$; alternative hypothesis) eli näytteessä on poikkeama jakaumasta (**efekti**). P-arvo taas kertoo sen, kuinka todennäköistä on se, että nollahypoteesi voitaisi todistaa oikeaksi tietyssä näytteessä. Mitä alempi p-arvo on, sen varmempi voi olla siitä, että nollahypoteesi on väärässä ja vaihtoehtoinen hypoteesi on oikeassa eli näytteessä on poikkeama jakaumasta. Lingvistissä piireissä p-arvon raja on yleensä 5 % ($p \leq 0.05$). Toisin sanottuna tämä tarkoittaa sitä, että nollahypoteesin todennäköisyys on ainoastaan 5 %. Vaihtoehtoisen hypoteesin todennäköisyys on siis 95 %. Huomioitavaa on se, että nollahypoteesi ei ole koskaan todistettavissa. Sen voi ainoastaan kumota tai olla kumoamatta eli voi sanoa, että ei ole todisteita nollahypoteesin kumoamiselle. (De Sutter 2018, 42-46.)

Tutkielmassa mainittiin aiemmin, että myös Chi-square testillä saadaan p-arvo (kts. 2.1.5). Tämä testi ei kuitenkaan kerro mitään poikkeaman suuruudesta tai suunnasta. Se antaa vain merkin siitä, missä

määrin näytteen havainnoidut arvot poikkeavat odotetusta teoreettisesta jakaumasta. Eli Chi-square testistä saatu p-arvo osoittaa sen, millä todennäköisyydellä saatu tulos on sattumaa. (Pandis 2016, 898.)

4.2.2 Metodit

Varsinaiset monimuuttujametodit, jotka valitsin varsinaista data-analyysiä varten ovat **binäärinen logistinen regressio**, **ehdollisen päättelyn puu eli päätös- tai luokittelupuu** (**Conditional Inference Tree (CIT), decision tree**) ja **satunnainen metsä (random forest).**

Logistinen regressio on eräänlainen laajennus suoraviivaisemmalle lineaariselle regressiolle. Lineaarista regressiota käytetään, kun vastemuuttujan oletetaan olevan jatkuva (Hosmer & Lemenshow 2004, 1). Jatkuva muuttuja tarkoittaa sitä, että muuttujan ”…kahden arvon välissä on ääretön määrä arvoja.” (KvantiMOTV). Logistinen regressio eroaa lineaarisesta sillä, että siinä vastemuuttuja on vähintään kaksijakoinen eli binäärinen (Hosmer & Lemenshow 2004, 1).

Logistisessa regressiossa mallinnetaan siis vastemuuttujan kahden tai useamman mahdollisen arvon ja kahden tai useamman selittävän muuttujan suhteita. Jos vastemuuttuja voi saada kaksi arvoa on malli binäärinen (binomiaali) ja jos taas mahdollisia arvoja on 3 tai useampi niin kyseessä on multinomiaali regressio. (Levshina 2015, 253.)

Binääristä logistista regressiota käytetään siis, kun tulokseksi saadaan kahtia jakautuva muuttuja. Siinä on tarkoituksena mallintaa kahtia jakautuva riippuva eli vastemuuttuja, käyttäen joukkoa selittäviä muuttujia. Tuloksena saadaan kokonaisindikaatio mallin globaaleista efekteistä ja kaikkien yksittäisten selittävien muuttujien merkittävyys, halliten samalla kaikkien muiden muuttujien vaikutusta tulokseen. Lisäksi saadaan tieto kaikkien selittävien muuttujien suhteellisesta vaikutuksesta vastemuuttujaan ja vaikutuksen suunnasta. (De Sutter 2018, 49-50.)

Esimerkki binäärisestä logistisesta regressiosta omasta aineistostani: minulla on aineistossani kaksijakoinen vastemuuttuja *Cond_RU,* joka kertoo, onko käännös tehty konditionaalilla vai ei. Tämä muuttuja voi siis saada kaksi eri arvoa, jotka ovat *Yes* (konditionaalilla) ja *No* (ei konditionaalilla). Binäärisen regression avulla mallinnan näiden kahden eri vastemuuttujan arvojen suhteita aineistoni selittäviin muuttujiin. Selittäviä muuttujia aineistossani voivat olla esim. kääntäjä, suomenkielisen konditionaaliesimerkin aikamuoto yms., jotka vaikuttavat siihen onko käännös tehty konditionaalilla vai ei.

Ehdollisen päättelyn puu on metodi, jota käytetään regressiossa ja binääriseen rekursiiviseen ositukseen (binary recursive partitioning) perustuvassa luokittelussa (Levshina 2015, 291). Binääristä rekursiivista ositusta käytetään luokitteluun, kun puun vastemuuttuja on *luokitteleva* (sillä on tietyt luokat, joihin se jakautuu), ja regressioon, kun puun vastemuuttuja on *jatkuva* (Merkle & Shaffer 2011, 161). Käyttämäni ehdollisen päättelyn puu on siis binääriseen rekursiiviseen ositukseen perustuvaa *luokittelua*, koska aineistoni vastemuuttuja on kategorinen (luokitteleva) muuttuja.

Binäärinen rekursiivinen ositus koostuu seuraavista toimenpiteistä: Algoritmi etsii selittävän muuttujan, joka liittyy vahvimmin vastemuuttujaan eli sen, jolla on pienin p-arvo. Seuraavaksi algoritmi päättää parhaan tavan osittaa datan kahteen osajoukkoon, jotka on jaettu vastemuuttujan eri arvoihin. Jos vastemuuttuja on kaksijakoinen (binäärinen), kuten tämän tutkielman aineistossa, niin yksi osajoukoista sisältää kaikki tapaukset, joissa vastemuuttuja saa arvon *A,* ja toinen osajoukko kaikki tapaukset, joissa vastemuuttuja saa arvon *B.* Viimeisenä toimenpiteenä kaikille mahdollisille osajoukoille toistetaan kaksi ensimmäistä toimenpidettä, kunnes tulokseksi ei saada enää yhtään p-arvoa, joka olisi yli 0.05 (eli $p < 0.05$). Ehdollisen päättelyn puun itse päättely (p-arvojen laskenta) perustuu permutaatioihin. Permutaatiot tarkoittavat tässä sitä, että havainnoitujen datapisteiden nimien paikkoja vaihdetaan useita kertoja, ja jokaisen paikan vaihdon jälkeen asiaankuuluvat testitilastot lasketaan. Tällä tavoin saadaan testitilastojen jakauma nollahypoteesin ollessa *ei eroa, ei liity toisiinsa* jne. (Levshina 2015, 291-292.)

Ehdollisen päättelyn puiden hyvä puoli on niiden helppo tulkinta, koska ne ovat visuaalisia ja niissä kaikkein merkittävin selittävä muuttuja on puussa ylimpänä. Huonoja puolia on taas se, että tulokset riippuvat melko vahvasti siitä tietystä näytteestä, jota käytetään. Tämä vaikeuttaa tulosten yleistämistä aineistoon, jota ei ole nähtävillä (otantakanta). On myös näytteitä siitä, että EPP on epäonnistunut havaitsemaan todella merkittäviä selittäviä muuttujia. EPP ei myöskään pysty käsittelemään satunnaisia tekijöitä. (De Sutter 2018, 55.)

Jotta pääsee yli EPP:n huonoista puolista, voi käyttää satunnaisten metsien (random forest) mallinnusta. Ehdollisen päättelyn puusta voi luoda satunnaisen metsän. Satunnaisten metsien perusidea on *kasvattaa* puumetsä (esim. 1500 puuta), joka perustuu selittävien muuttujien ja datapisteiden satunnaiseen valintaan. Satunnainen metsä luo siis olemassa olevan materiaalin pohjalta datamassiiveja, joissa tiettyjen muuttujien merkitystä vaihdetaan hieman suuntaan tai toiseen. Tämän jälkeen sulautetaan koko metsän puiden tulokset yhteen. Tämä sulautettu arvo kertoo jokaisen selittävän muuttujan vaikutuksesta vastemuuttujaan muiden selittävien muuttujien yhteydessä. Tästä saadaan tuloksena hyvin tarkka käsitys niistä selittävistä muuttujista, jotka vaikuttavat vastemuuttujaan. (Levshina 2015, 292.)

Tarkoituksenani on käyttää *binääristä logistista regressioanalyysiä* ja jatkotoimenpiteitä *ehdollisen päättelyn puut* sekä *satunnaiset metsät* aineistoni kaksijakoisen vastemuuttujan (Cond_RU) analysoimiseen. Ja selvittää siten, mitkä seikat vaikuttavat siihen, että suomen kielen konditionaalimuoto käännetään venäjään konditionaalilla.

4.3 Aineiston avaus

Microsoft Excel auttaa tietojen käsittelyssä ja niiden järjestelemisessä kvantitatiivista analyysiä varten. Solujen täyttämistä voidaan myös nopeuttaa erilaisilla ohjelman ominaisuuksilla. Sarakkeista voi esimerkiksi suodattaa tarvitsemansa asiat lisäämällä sarakkeille suodattimet (välilehdessä Data ja sieltä Filter, jos Excel on suomen kielellä niin Tiedot ja Suodata).

Otetaan esimerkki minun aineistostani. Suomenkielisessä konditionaaliesimerkissä esiintyy konditionaalin preesensmuoto verbistä tulla eli tulisi. Korpuksen hakutyökalu (Texthammer) on merkannut sen *_tulisi_* kontekstin keskellä, jotta sen löytäminen on helppoa. Haluan löytää kaikki konditionaaliesimerkit, jotka ovat tulla-verbin konditionaalin preesensmuotoja. Klikkaan Konteksti-sarakkeen suodatinta, etsin Text filters (tekstisuodattimet) osion ja valitsen sieltä Contains eli sisältää. Kirjoitan tähän *_tulisi_*, jolloin Excel suodattaa taulukon ja näyttää ainoastaan ne rivit eli esimerkit, joissa on tulla-verbin konditionaalin preesensmuoto.

Toinen ominaisuus, jota käytin paljon aineiston käsittelyssä on Replace eli Korvaa funktio (Ctrl+H). Sillä voi korvata tietyn aineistosta löytyvän asian suoraan kaikista soluista, joissa se esiintyy. Omassa aineistossani käytin tätä toimintoa esimerkiksi siinä, kun translitteroin kääntäjien nimiä.

Joissain tapauksissa korpuksessa oli rinnastettu ohi eli konditionaaliesimerkin käännösvastinetta ei ollut käännöksen kontekstisarakkeessa. Ja joissain tapauksissa konditionaaliesimerkki oli jätetty kääntämättä tai ikään kuin huomioimatta. Tämmöisissä tapauksissa jouduin poistamaan kyseiset konditionaaliesimerkit ja hakemaan niiden tilalle uusia esimerkkejä osakorpuksesta.

4.3.1 1000 esimerkkiä

Aineistoni koostuu siis 1000 konditionaalimuodon esimerkistä ja niiden käännösvastineista. Nämä 1000 esimerkkiä kopioitiin korpuksesta saaduista hakutuloksista Excel-taulukkoon. Excel on erittäin käytännöllinen aineiston koostamista ja sille tehtäviä jatkotoimenpiteitä varten, mutta mikä tahansa muu taulukkolaskentaohjelma käy myös.

Kuva 1. Ote hakutulosaineistosta Exceliin kopioituna

Context	Author	Title	Translation	Code	Translator	Title
Rane huokasi alistuneesti. Emme halunneet jatkaa keskustelua. Jaoksemme päällikkö, komisario Kalevi Kinnunen, oli alkoholisti. Piste. Minä olin hierarkiassa seuraava ja saisin ottaa jutun käsiini, kunnes Kinnunen _toipuisi_ humalastaan tai krapulastaan. Piste.	Lehtolainen Leena	Ensimmäinen murhani	Ране тяжело вздохнул. Разговор продолжать не хотелось. Руководитель нашего отдела комиссар Калеви Киннунен был алкоголиком. А я была руководителем следующего уровня и поэтому должна была принять удар на себя, пока Киннунен находился в запое или приходил в себя с похмелья.	lehtolainen_melnik_ensmurh_ru	Мельник, Татьяна	Мое первое убийство
Aluksi toivon, että _voisitte_ tunnistaa erään henkilön.	Nykänen, Harri	Ariel	Надеюсь, для начала вы сможете опознать одного человека.	Nyk_Pril_Ariel	Прилежаев, Иван	Ариэль
En uskonut serkun likaisia vihjeitä, en siihen että Rosalie _olisi_ tehnyt itselleen pahaa.	Oksanen Sofi	Kun kyyhkyset katosivat	Я не поверил грязным словам кузена, не могла Розали сделать с собой что-то дурное.	oksanen_sidorova_kyyhkyset_ru	Сидорова, Анна	Когда исчезли голуби
Majuri varoi hengittämästä poropoliisin suuntaan. Hurskainen _olisi_ tullut juovuksiin Remeksen vanhan viinan löyhkästä.	Paasilinna Arto	Hirtettyjen kettujen metsä	Майор остерегался дышать в сторону полицейского. Хурскайнен мог вполне захмелеть от запаха спиртного, который исходил от Ремеса.	paasilinna_zaikov_hirtettyjen_ru	Зайков, С. and Хотинский, Н.	Лес повешенных лисиц

Yllä olevan kuvan (kuva 1) valmiiksi korpuksesta saatujen sarakkeiden selitykset: Sarakkeessa *Context* on luonnollisesti se konteksti, jossa suomen kielen konditionaalimuoto esiintyy. Joko pelkkä lause, virke tai joissain tapauksissa kokonainen kappale, jonka sisällä konditionaaliesimerkki esiintyy. Toisessa sarakkeessa on korpuksen oma kontekstin koodi (Code), joka koostuu kirjailijasta_kääntäjästä_teoksesta_kielestä. Sarakkeessa *Author* on kirjailijan nimi ja *Title* sarakkeessa teoksen suomenkielinen nimi. Sarake *Translation* pitää sisällään venäjänkielisen käännöksen ja kontekstin, jossa se esiintyy. Toinen *Code* sarake on sama korpuksen sisäinen koodi, mutta se koskee luonnollisesti käännöstä. *Translator* sarakkeessa on kääntäjän nimi ja toisessa *Title* sarakkeessa teoksen käännetty venäjänkielinen nimi.

Aloittaessani aineistoni kvalitatiivisen käsittelyn, lisäsin taulukkoon seuraavat sarakkeet: **Fi_form**, **Fi_TL_AM**, **Fi_cndmerk**, **Fi_vclas**, **Ru_form**, **Ru_SL**, **Ru_TL**, **Ru_AM**, **Ru_ASP**, **Leks_ru**, **Cond_RU**.

Fi_form sisältää suomenkielisen konditionaaliesimerkin verbin perusmuodossa. Tarkistin sen jälkeen verbin tapaluokan ja aikamuodon, jotka laitoin sen sarakkeeseen **Fi_TL_AM**. Näistä luonnollisesti ainoastaan toinen eli aikamuoto vaihtui, sillä hain korpuksesta konditionaalimuotoja, joten kaikki esimerkit olivat tapaluokaltaan konditionaaliin kuuluvia. Kuten tutkielmassa aiemmin mainitsin (kts. 3.1) konditionaalilla on suomen kielessä vain kaksi tempusta: preesens ja perfekti. Merkitsin siis taulukkoon aikamuodosta riippuen joko cnd_pres tai cnd_perf.

Fi_cndmerk sarakkeeseen merkitsin suomenkielisen konditionaalimuodon merkityksen. Määritelmät ja taulukkooni tekemät merkinnät ovat seuraavat: Intensionaalinen asiaintila (yhtenä ajateltavissa olevista vaihtoehdoista eli mahdollinen) **mahd.** Tahtotulkintainen konditionaali (vetoomus, ehdotus, kutsu tai tiedustelu) **tht.** Konditionaalin todenvastaisuus ja toteutumattomuus

(esim. retoriset kysymykset) **tvast.** Ehtotulkintainen (jos jotain tapahtuu, niin sitten jotain muuta tapahtuu) **eht.** Merkitysten määritelmät katsoin Isosta suomen kieliopista (VISK § 1592-1595). Päätin jakaa suomenkielisten verbiesimerkit myös niiden semanttisen merkityksen mukaan **Fi_vclas** sarakkeeseen. Määritelmät näille katsoin myös Isosta suomen kieliopista (VISK § 445) ja ne löytyvät selitettynä tutkielman sivulta 39.

Etsin sen jälkeen venäjänkielisen käännöksen sarakkeesta käännösvastineen suomenkieliselle konditionaalimuodolle ja laitoin sen sarakkeeseen **Ru_form**. Translitteroin nämä käyttäen ISO 9 -formaatin translitterointia. Tähän löytyy useampiakin nettisivuja, joilla on automaattinen translitterointijärjestelmä. Itse käytin sivustoa *translitteration.com*. Tämän jälkeen tarkistin, minkä sanaluokan edustaja käännösvastine on ja merkitsin sanaluokan sarakkeeseen **Ru_SL**. Käännösvastineen tapaluokan merkitsin **Ru_TL** sarakkeeseen. Seuraavaksi merkitsin sen aikamuodon **Ru_AM** ja aspektin **Ru_ASP** omiin sarakkeisiinsa. **Leks_ru** -sarake on niitä tapauksia varten, joissa venäjänkielisessä käännöksessä on konditionaali ilmaistu jollain muulla kuin itse konditionaalimuodolla. Tämä ei siis käsitä kaikkia niitä tapauksia, joissa käännös on tehty muulla tapaa kuin venäjän kielen konditionaalia käyttäen vaan sellaisia, joissa esimerkiksi ehdollisuus on selvästi ilmaistuna, käyttäen jotain muuta leksikaalista keinoa kuten venäjän kielen partikkelia *li*. Esimerkki omasta aineistostani: Lehtolainen, Leena, Kuparisydän: ”*Pitäisikö tarkistaa*, jos se kuitenkin olisi hengissä?”. Käännös: Džafarova-Viitala, Tais′â: ”*Не проверить ли*, может, он еще жив?”.

Viimeisenä sarakkeena lisäsin sarakkeen **Cond_RU**, johon on merkittynä, onko käännösvastineen tapaluokka (Ru_TL sarake) konditionaali (Yes) vai ei (No). Merkitsin kaikkiin taulukon tyhjiin ruutuihin **X**. Tämmöisiä ovat esimerkiksi konditionaaliesimerkit, joissa käännösvastine on konditionaalissa, jolloin laitoin käännösvastineen aikamuodon (Ru_AM) sarakkeeseen **X**, koska venäjän kielen konditionaalilla ei ole aikamuotoa. Tämä selkeyttää aineiston jatkokäsittelyä, koska silloin siellä ei ole tyhjiä ruutuja, jotka saattavat vaikeuttaa aineiston käsittelyä ohjelmassa R.

4.3.2 Aineiston tilastoja: esimerkit

Varsinaisen materiaalin analysoinnin ja käsittelyn tein Microsoft Excelillä ja ohjelmalla R sekä sen ohjelmistoympäristöllä RStudio. Heti alussa ilmeni kuitenkin yksi ongelma: kyrilliset aakkoset eivät näy oikein R:ssä. Tämä ongelma johtuu siitä, että R:llä on ongelmia muiden kuin latinalaisten aakkosten käsittelemisessä Windowsissa. Joten jouduin R:ssä käsittelyä varten käymään läpi aineistoni uudestaan ja translitteroimaan kaikki venäjänkieliset käännökset, kääntäjien nimet ja teosten venäjänkieliset nimet. Käytin translitterointiin jälleen samaa translitteration.com sivustoa ja

ISO 9 järjestelmää. Saatuani aineistoni RStudioon huomasin, että siinä oli tutkimukseni kannalta tarpeettomia sarakkeita. Poistin siis RStudiossa aineistosta seuraavat sarakkeet Konteksti, Koodi_fi, Käännös, Koodi_ru ja Teos_ru. Tämän jälkeen aineistoon jäi 12 saraketta eli muuttujaa ja 1000 riviä eli havaintoa (arvoa) niistä.

Vaikka aineiston haku tehtiin korpuksessa niin, että saatu näyte olisi satunnaisessa järjestyksessä otettu, on siinä silti tiettyjen muuttujien arvoissa selviä trendejä. Esimerkiksi kirjailijoissa, joiden teoksista otetuista konditionaalin esimerkeistä näyte koostuu, on selviä eroja siinä, kenen teoksista on otettu mikäkin määrä esimerkkejä. Näistä Sofi Oksasen teoksista on 394 esimerkkiä eli 39 prosenttia näytteestä ja Leena Lehtolaisen teoksista on 244 esimerkkiä eli 24 prosenttia koko näytteestä. Muiden 10 kirjailijan teoksia on 7 ja 1 prosentin väliltä. Jakauma ei siis ole ehkä paras mahdollinen kirjailijoiden osalta, sillä n. 64 prosenttia konditionaaliesimerkeistä on otettu kahden kirjailijan teoksista ja vain 36 prosenttia esimerkeistä muilta kirjailijoilta.

Taulukosta 1 voi nähdä, että aineistoni esimerkkien määrä ja kirjailijoiden teosten sanamäärä ParFin osakorpuksessa vastaavat melko hyvin toisiaan. Oikeastaan ainoastaan Sofi Oksasen kohdalla on huomattava poikkeama. Aineistoni esimerkeistä 39,4 prosenttia on Oksasen teoksista, mutta korpuksen sanamäärästä Oksasen teokset ovat vain 20,5 prosenttia, eli noin puolet vähemmän. Joten muiden kirjailijoiden kuin Oksasen osalta voisi sanoa jakauman olevan korpusta vastaava.

Tämä ei ole tutkimukseni kannalta kovinkaan tärkeä seikka, koska sillä ei ole merkitystä, mistä konditionaalin käytön esimerkit on otettu, koska tutkin sitä, miten ne käännetään suomesta venäjään. Toki, jos kaikki esimerkit olisi otettu yhdestä teoksesta niin myös kääntäjiä olisi ollut vain yksi, mikä taas vaikuttaisi tutkimukseeni.

Taulukko 1. Kirjailijajakauma näytteessä & sanamäärä ParFin osakorpuksessa

Kirjailija	Esimerkkejä	% esimerkeistä	Sanamäärä ParFin	% sanamäärästä
Haahtela, Joel	37	3,7	24938	3,8
Hotakainen, Kari	44	4,4	47378	7,3
Konkka, Anita	13	1,3	22088	3,4
Krohn, Leena	33	3,3	34082	5,3
Lehtolainen, Leena	244	24,4	148993	23,2
Mäkelä, Hannu	33	3,3	31806	5
Nykänen, Harri	44	4,4	46168	7,1
Oksanen, Sofi	394	39,4	131732	20,5
Paasilinna, Arto	69	6,9	41784	6,5
Rimminen, Mikko	48	4,8	57107	8,9
Salminen, Arto	19	1,9	23246	3,6
Sinisalo, Johanna	22	2,2	32824	5,1
Yhteensä	**1000**	**100 %**	**642146**	**100 %**

Luonnollisesti myös tietyt kääntäjät ovat yliedustettuina aineistossa, koska tämä määrä korreloi tietenkin sen kanssa, kenen teoksista esimerkkejä on otettu ja kuinka monta kappaletta. Esimerkiksi Tais′â Džafarova-Viitala sekä Anna Sidorova ovat selvästi eniten käännöksiä aineistossa tehneet kääntäjät (Taulukko 2). Tällä taas, toisin kuin kirjailijoiden yliedustuksella, on merkitystä omassa tutkimuksessani, koska siihen, onko käännös tehty venäjän kielen konditionaalilla vai ei, saattaa vaikuttaa myös se, että kuka käännöksen on tehnyt. Tässä taulukko kääntäjistä ja kirjailijoista, joiden teoksia he ovat kääntäneet ja esimerkkien määrästä:

Taulukko 2. Kääntäjät & Kirjailijat

Kääntäjä - Kirjailija	**Esimerkkejä**
Džafarova-Viitala, Tais′â	**265**
Konkka Anita	13
Lehtolainen Leena	82
Oksanen Sofi	170
Ioffe, Èleonora	**33**
Mäkelä Hannu	33
Mel′nik, Tat′âna	**162**
Lehtolainen Leena	162
Olykajnen, Leonid	**19**
Salminen Arto	19
Priležaev, Ivan	**81**
Haahtela Joel	37
Nykänen, Harri	44
Sidorova, Anna	**297**
Krohn Leena	25
Oksanen Sofi	224
Rimminen Mikko	48
Tinovickaâ, Evgeniâ	**8**
Krohn Leena	8
Ureckij, Il′â	**44**
Hotakainen Kari	44
Virolajnen, Laura and Ioffe, Èleonora	**22**
Sinisalo Johanna	22
Zajkov, S. and Hotinskij, N.	**69**
Paasilinna Arto	69
Yhteensä	**1000**

Suomenkielisten esimerkkien konditionaalimuodon muodostavista verbeistä hieman tilastoja. Jätän tässä mainitsematta kaikki verbit, joita esiintyi alle 10 esimerkin verran. Näitä vähintään 10 kertaa esiintyviä oli 17 kappaletta (Taulukko 3). Verbit *ehtiä, haluta, haluttaa, käydä, lähteä, löytää, mennä, ottaa, pystyä, tarvita, tehdä, tietää* esiintyvät kaikki 10-15 kertaa aineistossa ja yhteensä 145 kertaa. Verbi *tulla* esiintyy aineistossa 44 kertaa, *saada* 47 kertaa ja verbi *pitää* 57 kertaa. *Voida* verbi on aineistossa 99 kertaa ja *olla* verbi 158 kertaa.

Taulukko 3. Verbiesiintymät

Verbi	Freq	Verbi	Freq
olla	158	lähteä	12
voida	99	mennä	12
pitää	57	pystyä	11
saada	47	ehtiä	10
tulla	44	haluttaa	10
haluta	15	käydä	10
tarvita	15	löytää	10
tehdä	15	ottaa	10
tietää	15		
		Yhteensä	**550**

Aineistoa käsitellessä lisäsin useille näistä verbeistä Fi_form sarakkeeseen selittävän lisäsanan sulkuihin. Tämän tarkoituksena oli lähinnä helpottaa aineiston tarkastelua ja tarkistusta, koska tällöin minun ei tarvinnut miettiä ja tarkistaa asiaa Konteksti ja Käännös-sarakkeista. Otetaan yksi esimerkki: suomenkielisessä konditionaaliesimerkissä on muoto ”…se _olisi_ ollut matkalla ylöspäin…” (eli konditionaalin perfektimuoto verbistä *olla*) ja tämä on käännetty venäjään verbillä *dvigat'sâ* (двигаться) eli 'liikkua'. Tässä tapauksessa laitoin Fi_form sarakkeeseen *olla (matkalla)*. Tilastot, jotka tässä verbiesiintymistä kerron, ovat kuitenkin sellaiset, joissa olen jättänyt nämä suluissa olevat lisämerkinnät huomioimatta. Esimerkiksi *olla* verbiä esiintyy aineistossa sellaisenaan 128 kertaa ja erilaisilla lisämerkinnöillä olevia versioita on 30 (erilaisia lisäsanoja on 24, muutama toistuu kaksi tai kolme kertaa), mutta kuitenkin tilastoissa laskin olla-verbin käyttömääräksi 158.

Nämä 17 verbiä, jotka listasin, kattavat yli puolet (550) kaikista esimerkeistä. Erilaisia verbejä on aineistossa yhteensä 278 kappaletta. Tästä johtuen yksittäisten verbiesiintymien keskiarvo on 3,59. Esiintymien mediaani on kuitenkin 1, mikä selittyy sillä, aineiston 278:sta erilaisesta verbistä 184 (66 %) on aineistossa vain yhden kerran. Esiintymien keskihajonta on 11,65 eli keskimääräinen poikkeama keskiarvosta on melko suuri (8,05). Tämä taas selittyy sillä, että listaamistani 17:sta verbistä koostuu 55 prosenttia aineiston esimerkeistä, mikä tarkoittaa luonnollisesti sitä, että loput 45 prosenttia esimerkeistä koostuu 261:stä eri verbistä.

Seuraavaksi tilastoja itse merkitsemistäni VISK:n määritelmien mukaisista konditionaalimerkityksistä (Fi_cndmerk) ja esimerkit niistä omasta aineistostani. Konditionaalimerkitykset, joihin jaoin aineistoni esimerkit ovat seuraavat: Intensionaalinen asiaintila **mahd** (esimerkki 1), tahtotulkintainen konditionaali **tht** (2), konditionaalin todenvastaisuus ja toteutumattomuus **tvast** (3) sekä Ehtotulkintainen konditionaali **eht** (4).

(1) Hän **_olisi_ voinut** polttaa kaiken pikku hiljaa, käyttää kirjat sytykkeiksi, mutta tuntui tärkeältä päästä niistä eroon heti. (Sofi Oksanen, Puhdistus)

(2) Mä en **_haluaisi_** sekoittaa isääni ja Muurialaa tähän. (Leena lehtolainen, Ensimmäinen murhani)

(3) …ja jos olisin kiinnostunut, hän osaisi myös neuvoa paikan, jossa ei **_tarvitsisi_** olla iltaisin yksin, you know what I mean, hän lisäsi vielä englanniksi. (Joel Haahtela, Perhoskerääjä)

(4) En **_olisi_ muistanut** miehen nimeä, ellei hän olisi sanonut sitä, mutta muistin kasvot.) Nämä eivät myöskään jakaudu tasaisesti:

Taulukko 4. Konditionaalimerkitysjakauma

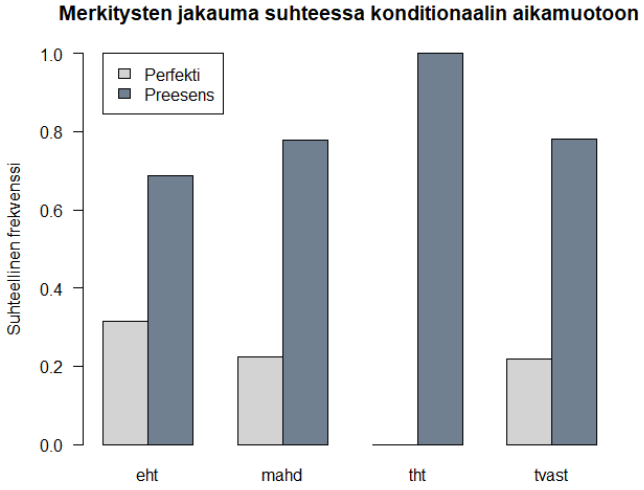
Fi_cndmerk	Kpl
eht	153
mahd	193
tht	48
tvast	606
Yhteensä	**1000**

Aineiston tämän tyyppisten luokittelujen merkitseminen on vaikeaa ja vaatii valmistelua sekä harjoittelua, koska esimerkiksi konditionaalimuotojen merkityksiä merkatessa on aina olemassa joitain marginaalitapauksia, jotka voidaan tulkita eri tavoin. Suurissa projekteissa voi olla käytettävissä useita aineiston käsittelijöitä, joiden tekemiä merkintöjä voidaan vertailla keskenään. Tutkielmani tapauksessa tämä ei ole mahdollista, sillä merkinnät on tehnyt yksi käsittelijä eli tutkielman kirjoittaja. Tästä syystä on mahdollista, että merkinnöissä on virheitä, mutta tätä on yritetty kompensoida suurella esimerkkimäärällä.

Aineistoni suomenkielisten konditionaaliesimerkkien aikamuotoa tarkastellessa huomaa myös, että niissäkään ei ole satunnaisen aineiston valinnan syystä tai ansiosta, tullut tasaista jakaumaa aikamuodoissa (kts. Kuva 2). Konditionaalin preesensmuodon edustajia on esimerkeistä 777 kappaletta ja konditionaalin perfektimuotoa esiintyy 223 kertaa. Eli esimerkeistä noin ¾ on preesensissä ja ¼ on perfektissä. Tämä näkyy luonnollisesti myös, kun tarkastellaan konditionaalimerkitysten jakaumaa suhteessa konditionaalin aikamuotoon. Ehtotulkintaisen konditionaalimuodon 153:sta esimerkistä 105 (68 %) esimerkkiä on preesensissä ja 48 (32 %) perfektissä. Intensionaalista asiantilaa merkitsevän konditionaalimuodon 193:sta esimerkistä 150 (78 %) on preesensissä ja 43 (22 %) perfektissä. Tahtotulkintaisen konditionaalimuodon merkityksen esimerkkejä oli aineistossani vähiten eli 48 esimerkkiä (eli vain 4,8 % näytteestä), mutta niistä kaikki 48 esimerkkiä ovat aikamuodoltaan preesensissä. Suurin osa konditionaaliesimerkeistä ovat todenvastaisuutta ja toteutumattomuutta edustavia konditionaaliesimerkkejä. Niitä on aineistossa kokonaisuudessaan 604 kappaletta ja niistä 474 (78 %) esimerkkiä ovat preesensissä ja 130 (22 %)

perfektissä. Tätä tarkastellessa huomaa, että muiden kuin tahtotulkintaisen konditionaalimuodon merkityksen esimerkeissä jakauma vastaa melko läheisesti yleistä aikamuotojen jakaumaa.

Kuva 2. Konditionaalimerkitykset jaettuna aikamuotoihin



Suomenkielisten konditionaaliesimerkkien viimeisenä merkintäsarakkeena on **Fi_vclas** eli verbien luokittelu semanttisen merkityksen mukaan (Taulukko 5). Verbien jako oli seuraava: olemista tai oleskelua ilmaisevat (OLE), liikettä ilmaisevat (LII), tekemistä ilmaisevat (TEK), tunnetta ilmaisevat (TUN), havaintoa ilmaisevat (HAV), tiedontilaa ilmaisevat (TIE), viestintää ilmaisevat (VIE), modaaliset verbit (MOD), asioiden suhteita ilmaisevat verbit (SUH) ja kaikki muut verbit (MUU). Merkitsin ne taulukkoon siis näillä suluissa olevilla kolmikirjaimisilla lyhenteillä ja yhteensä semanttisen merkityksen kategorioita oli 10. Tämän jaon tarkoituksena on se, että tutkin, vaikuttaako suomenkielisen verbin semanttinen merkitys siihen, millä tavoin se on käännetty venäjään.

Taulukko 5. Suomenkielisten verbien jakauma semanttisen merkityksen mukaan

Fi_vclas	Freq
MOD	247
OLE	201
TEK	200
LII	128
HAV	50
TUN	47
MUU	44
TIE	37
VIE	34
SUH	12
Yhteensä	**1000**

4.3.3 Aineiston tilastoja: käännökset

Venäjänkielisen käännösvastineen (Ru_form) sanaluokkia (Ru_SL) oli yhteensä 9 erilaista. Näistä huomattavasti suurin joukko eli 887 esimerkkiä, eli 88,7 prosenttia koko näytteestä, käännettiin verbillä. Esimerkeistä 60 kappaletta käännettiin predikatiivilla, 20 kappaletta adjektiivilla, 14 kappaletta partisiipilla, 8 esimerkkiä gerundilla, 6 kappaletta fraasilla, 3 kappaletta adverbilla ja substantiivilla sekä partikkelilla käännettiin molemmilla ainoastaan 1 esimerkki. Aineiston kvantitatiivisen analyysin helpottamista varten erotin sellaiset tapaukset, jotka liittyvät kielioppiin eli verbi, predikatiivi ja gerundi, niistä, jotka syntyivät parafrasoinnin ja muiden muutosten seurauksena. Nämä loput sanaluokat, joilla käännös on tehty, ovat kategoriassa *Muut* (Taulukko 6).

Taulukko 6. Käännösvastineiden sanaluokkajakauma

Ru_SL	Käännösvastineita
Verbi	887
Predikatiivi	60
Gerundi	8
Muut	45
Yhteensä	**1000**

Venäjänkielisten käännösvastineiden tapaluokkajakauma on seuraavanlainen (kts. Taulukko 7): Konditionaalilla käännettiin vain 286 esimerkkiä ja indikatiivilla yli 2 kertaa enemmän eli 607 esimerkkiä. Imperatiivilla käännettiin ainoastaan 4 esimerkkiä, joista kaksi on olla-verbin eli *byt'* (быть) kääntämisessä käytettyjä. Toinen käännöksistä aineistossani (esimerkki 5). Yhtä imperatiiveista käytettiin sanoa-verbin käännöksessä, jossa se kääntyy verbillä *predstavit'sâ* (6) Ja viimeistä käytettiin käännöksessä, jossa käännettiin naida-verbi eli *ženit'sâ* (7). Näistä neljästä imperatiivilla käännetystä esimerkistä 3 käännöstä teki Prileževaev, Ivan, joka käänsi lisäksi toisen olla-verbin käännöksistä.

(5) Kaikilla niillä oli vaikka **_olisi_ ollut** keskikesä. (A. Salminen, Ei-kuori) / У них у всех такие, **будь** хоть середина лета. (Spasibo net, kääntäjä L. Olykajnen)

(6) En olisi muistanut miehen nimeä, ellei hän **_olisi_ sanonut** sitä, mutta muistin kasvot. (H. Nykänen, Ariel) / Не **представься** он, я не вспомнил бы его имени, но лицо было знакомо. (Ariel′, kääntäjä I. Prileževaev)

(7) Tämä kaikki olisi nyt sinun, jos **_olisit_ nainut** Karmelan. (H. Nykänen, Ariel) / Все это принадлежало бы сейчас тебе, **женись** ты на Кармеле. (Ariel′, kääntäjä I. Prileževaev)

Iso X-kirjain tarkoittaa sitä, että esimerkin käännöksellä ei ole tapaluokkaa, mikä tarkoittaa siis tietysti sitä, että siihen kuuluvat ne 103 esimerkkiä, joissa käännös on tehty jollain muulla kuin verbillä. Tähän kategoriaan eivät kuitenkaan sisälly sellaiset käännökset, jotka kääntyvät venäjän kielen konditionaalin predikatiivimuodolla eli esimerkiksi sellaiset käännökset, joissa on käytetty muotoa *možno bylo by* (esimerkki 8). Tästä johtuu siis 10 käännösvastineen poikkeama verbimäärän ja X-merkittyjen tapaluokattomien käännösten välillä (1000 esimerkkiä – 887 verbiä = 113 tapaluokatonta käännöstä).

(8) Kreelin Marialta **_voisi_** tietenkin kysyä tai vaikka viedä mytyn hänelle… (S. Oksanen, Puhdistus) / У Марии Крел **можно было бы**, конечно, спросить или отнести к ней узелок… (Očiŝenie, käänt. T. Džafarova-Viitala)

Taulukko 7. Käännösvastineiden tapaluokkajakauma

Ru_TL	Freq	%
cnd	286	28,6
imper	4	0,4
ind	607	60,7
X	103	10,3
Yhteensä	**1000**	**100**

Käännösvastineita (Ru_form) on aineistossa yhteensä 476 erilaista. Yksittäisen käännösvastineen esiintymien keskiarvo on 2,10, mutta esiintymien mediaani on 1. Tämä selittyy sillä, että erilaisista käännösvastineista 360 on sellaisia, jotka esiintyvät aineistossa vain kerran. Esiintymien määrän keskihajonta on kuitenkin vain 5,99, lähes puolet pienempi kuin suomenkielisten esimerkkien verbien keskihajonta. Tämä ei ole kuitenkaan täysin verrattavissa oleva tilasto, sillä käännösvastineet eivät koostu 100 prosenttisesti verbeistä, kuten suomenkieliset esimerkit.

Adjektiiveista esiintyy käännöksissä adjektiivi *dolžen* (esimerkki 9) eniten eli 7 kertaa 20:stä vastineesta. Adverbeissa, fraaseissa, gerundeissa ja partisiipeissa ovat kaikki yksittäiset vastineet erilaisia. Luonnollisesti myös substantiiveissa ja partikkeleissa ovat niiden yksittäiset vastineet erilaisia, koska niitä on vain yksi kutakin. Esimerkki gerundin käytöstä aineistoni käännöksissä (10).

(9) Huuhkajan **_pitäisi_** ottaa avain ja lentää sitten karkuun! Tämä kykenisi siihen! (H. Mäkelä, Pekka Peloton) / Филин **должен** схватить ключ и улететь! Ведь это-то он может! (Besstrašnyj Pekka, käänt. È. Ioffe)

(10) Enkeli sulkee silmänsä hitaasti niin kuin **_koettaisi_** kaikin voimin hillitä itseään. (J. Sinisalo, Ennen päivänlaskua ei voi) / Ангел медленно закрывает глаза,

словно **старaясь** восстановить душевное равновесие. (Troll′ käänt. L. Virolajnen & È. Ioffe)

Predikatiivia on käytetty 62:ssa käännöksessä ja yhteensä 18 eri predikatiivia. Näistä 62:sta 55 kappaletta ei ole konditionaalissa eli ainoastaan 7 on konditionaalissa. Predikatiiveista kaksi predikatiivia muodostavat 46 vastinetta 62:stä ja lopuista 16 predikatiivivastineesta yksikään ei esiinny yli viittä kertaa. Nämä kaksi eniten aineistossa esiintyvää predikatiivivastinetta ovat *možno* (esimerkki 11), jolla on 24 esiintymää ja *nado* (12), joka esiintyy aineistossa 22 kertaa.

(11) Kiristin hinnan 687 000 markkaan, sen verran yli he **_joutuisivat_** onnestaan maksamaan. (K. Hotakainen, Juoksuhaudantie) / Я загнул 687 тысяч. За счастье, друзья, **можно** и переплатить. (Ulica okopnaâ, käänt. I. Ureckij)

(12) Annamari oli todennut Armin muistotilaisuudessa, että Sannan nimi **_pitäisi_** viimeinkin kaiverruttaa hautakiveen. (L. Lehtolainen, Harmin paikka) / На поминках Арми Аннамари сказала, что **надо** выбить имя Санны на могильном камне. (Zmei v raû, käänt. T. Mel′nik)

Verbeistä esittelen tässä ainoastaan ne, jotka esiintyvät 10 kertaa tai useammin (kts. Taulukko 8). Näitä on 9 kappaletta. Ei ole tietenkään yllättävää, että näissä toistuvat samat verbit, kuin suomenkielisissä esimerkeissä. Ylivoimaisesti eniten aineiston käännösvastineissa esiintyy *olla*-verbi *byt′* (быть), joka esiintyy 105 kertaa. Tämä on tietenkin odotettavissa, sillä myös suomenkielisissä konditionaaliesimerkeissä olla-verbi esiintyy eniten. Näistä esiintymistä 34 vastinetta on käännetty konditionaalilla, 69 kappaletta indikatiivilla ja 2 imperatiivilla. Seuraavaksi eniten käännösvastineissa, samoin kuin esimerkeissä, on voida-verbiä eli imperfektiaspektista muotoa *moč′* (мочь) ja sen perfektiaspektista muotoa *smoč′* (смочь). Näistä *moč′* esiintyy aineistossa yhteensä 66 kertaa. Niistä 21 on konditionaalilla ja 45 indikatiivilla käännettyjä. Perfektiaspektista *smoč′* on huomattavasti vähemmän, vain 21 kappaletta, ja sen esiintymistä 2 on käännetty konditionaalilla ja 19 indikatiivilla.

Verbi *haluta/tahtoa* eli *hotet′* (хотеть) esiintyy käännösvastineissa 15 kertaa, joista konditionaalilla on käännetty 6 ja indikatiivilla 9 kappaletta. Verbi *täytyä/joutua* eli *prijtis′* (прийтись) on aineistossa 13 kertaa. Sen esiintymistä 2 on käännetty konditionaalilla ja 11 indikatiivilla. *Tulla*-verbi *prijti* (прийти) ja *tietää*-verbi *znat′* (знать) esiintyvät molemmat aineistossa 12 kertaa. Niiden molempien tapaluokkajakauma on identtinen eli 4 käännöstä konditionaalilla ja 8 indikatiivilla. Verbi *stat′* (стать) eli *tulla* joksikin (*muuttua, valmistua* joksikin) esiintyy aineistossa 11 kertaa. Näistä

ainoastaan yksi on käännetty konditionaalilla ja 10 indikatiivilla. Verbi *stoit'* (стоить) eli *maksaa* tai *kannattaa* esiintyi aineistossa 10 kertaa ja niistä 6 on konditionaalilla ja 4 indikatiivilla käännettyjä.

Taulukko 8. Aineistoni 9 eniten esiintyvää verbiä

Verbi	Freq	%
byt'	105	39,6
moc'	66	24,9
smoc'	21	7,9
hotet'	15	5,6
prijtis'	13	4,9
prijti	12	4,5
znat'	12	4,5
stat'	11	4,1
stoit'	10	3,7
Yhteensä	**265**	**100**

Aineistoni verbit vastaavat hyvin taulukkoa (Taulukko 9) yleisimmin venäjän kielen konditionaalimuodon kanssa esiintyvistä verbeistä (Dobrušina 2014, 107). Tässä taulukossa on koottu aineisto osakorpuksesta, joka on 1970 vuodesta eteenpäin, ja oma aineistoni on 1976 vuodesta eteenpäin. Taulukossa tarkastellaan toki vain verbejä, jotka esiintyvät muodossa: perfekti + *by* partikkeli. Oma aineistoni vastaa kuitenkin tuota taulukkoa hyvin. Siinä reilusti yleisimmin esiintyvät verbit ovat olla-verbi *byt'* ja voida-verbi *moc'*, joita seuraa verbi *hotet'* eli haluta. Toisin kuin omassa aineistossani, verbi *smoc'*, eli perfektiaspektinen muoto verbistä *moc'*, esiintyi Dobrušinan taulukon aineistossa huomattavasti vähemmän, noin 3 kertaa vähemmän kuin *hotet'*. Mielenkiintoista on huomata, että myös verbien suhteelliset frekvenssit ovat melko samanlaiset molemmissa aineistoissa.

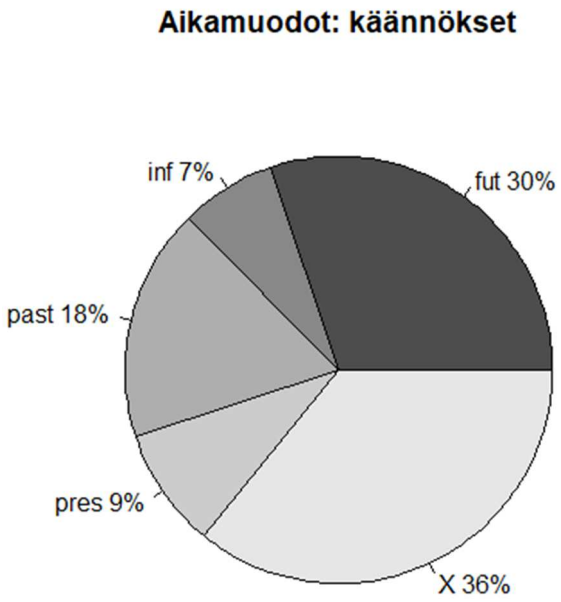
Otin tässä vertailua varten Dobrušinan taulukosta saman määrän eniten esiintyviä verbejä eli 9 eri verbiä. Omassa aineistossani yhdeksän eniten esiintyvää verbiä ovat aineistossa yhteensä 265 kertaa (Taulukko 8). Näistä yksittäisistä esiintymistä 39,6 prosenttia (105 kpl) on *byt'* verbiä. Dobrušinan taulukon 51815:stä esiintymästä *byt'* verbiä on 41,5 prosenttia (21543 kpl). Taulukkoni verbiesimerkeistä 24,9 prosenttia ja Dobrušinan taulukossa 24,2 prosenttia koostuu verbistä *moc'*. Lisäksi verbi *byt'* esiintyy tutkielmani aineistossa 1,59 kertaa enemmän kuin verbi *moc'* ja Dobrušinan taulukossa *byt'* esiintyy 1,71 kertaa enemmän kuin *moc'*. Näistä seikoista voisi päätellä, että näyte, josta aineistoni koostuu, on tutkimusta ajatellen, ainakin venäjänkielisten verbien osalta, hyvä näyte.

Taulukko 9. Yleisimmin konditionaalimuodossa esiintyvät verbit (Dobrušina)

Verbi	Freq	%
byt′	21543	41,5
moč′	12540	24,2
hotet′	4370	8,4
kazat′sâ	4077	7,8
hotet′sâ	3500	6,7
stat′	2008	3,8
smoč′	1567	3,0
skazat′	1196	2,3
sledovat′	1014	1,9
Yhteensä	**51815**	**100**

Aineistoni käännösvastineiden aikamuotojakauma on seuraavanlainen:

Kuva 3. Käännösvastineiden aikamuotojakauma



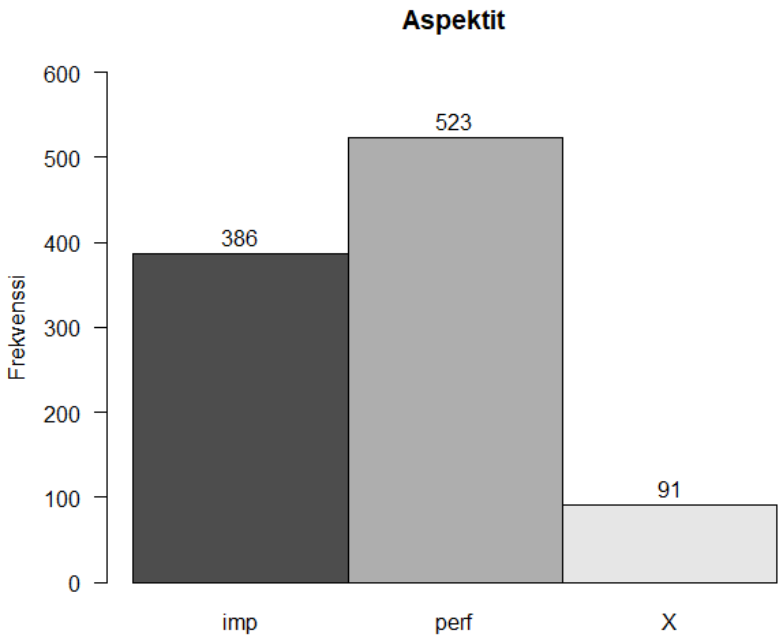
Kuten ympyrädiagrammista (Kuva 3) näkyy, aineistoni käännösvastineista 64 prosenttia omaavat aikamuodon ja 36 prosenttia on aikamuodottomia (**X**). Aikamuodottomiin kuuluvat kaikki tapaluokaltaan konditionaalia ja imperatiivia edustavat käännösvastineet, koska niillä ei ole aikamuotoa venäjän kielessä, sekä tapaluokattomat käännösvastineet eli kaikki muut paitsi verbit (adjektiivit, adverbit jne.). Myös suurin osa predikatiivilla käännetyistä muodoista eli 40/60 ovat aikamuodottomia.

Futuurissa (**fut**) on 30 prosenttia aineiston käännösvastineista eli lähes puolet aikamuodollisista käännösvastineista (30/64 %). Tähän kuuluvat verbit ja predikatiivit, jotka ovat futuurissa. Seuraavaksi suurin määrä käännösvastineita eli 18 prosenttia on perfektissä (**past**). Tähän kuuluvat

perfektissä olevat verbillä, predikatiivilla ja partisiipilla kääntyvät esimerkit. Preesensissä (**pres**) olevia käännösvastineita on 9 prosenttia. Niihin kuuluvat preesensissä olevat verbillä, predikatiivilla ja partisiipilla kääntyvät vastineet. Infinitiivissä (**inf**) olevia käännösvastineita on vähiten eli 7 prosenttia. Nämä koostuvat verbeistä, jotka ovat indikatiivissa ja infinitiivissä.

Aineistoni käännösvastineiden aspektijakauma (Kuva 4) on hyvin perfektivoittoinen sillä yli puolet, eli 523 käännösvastinetta, ovat aspektiltaan perfektiivisiä. Näihin kuuluvat perfektiaspektin omaavat verbit, gerundit ja partisiipit. Imperfektiaspektin omaavia käännösvastineita on vähemmän eli 386 kappaletta. Nämä koostuvat verbeistä, gerundeista ja partisiipeista, jotka ovat imperfektiaspektissa. Aspektittomia käännösvastineita on aineistossani 91 kappaletta. Näihin kuuluvat kaikki muut sanaluokat eli adjektiivit, adverbit, fraasit, partikkelit, predikatiivit ja substantiivit.

Kuva 4. Käännösvastineiden aspektijakauma



Aineiston käännösvastineet jaettiin vielä jatkotutkimusta varten konditionaalilla ja muilla muodoilla käännettyihin. Mielenkiintoista on, että käännösvastineista vain 286 eli vain 28,6 prosenttia on konditionaalilla käännettyjä. Analyysiin käytetyssä binäärisessä logistisessa regressiossa tätä saraketta käytetään vastemuuttujana.

Taulukko 10. Onko käännösmuoto konditionaalissa?

Cond_RU	Freq	%
No	714	71,4
Yes	286	28,6
Yhteensä	**1000**	**100**

Viimeisenä ja oikeastaan vähäisimpänä tilastona on leksikaaliset keinot, joilla on joissain tapauksissa ilmaistu konditionaalia sellaisissa tapauksissa, joissa käännöstä ei ole kuitenkaan tehty konditionaalilla. Vähäisimpänä siksi, että tämä tilasto on täysin itseni tekemä, eli oman harkintani varassa tehty, eikä mistään tarkistettu tai varmistettu. Suurin osa, eli 865 käännösmuotoa, eivät omaa mitään leksikaalista keinoa, jolla konditionaali on ilmaistu. Tähän joukkoon kuuluvat luonnollisesti myös ne käännösmuodot, jotka ovat konditionaalissa. Loput 135 leksikaalisen keinon omaavaa käännösmuotoa koostuvat verbeistä, partikkeleista, sidesanoista, adjektiiveista, adverbeista.

Näiden aineistoni tilastojen perusteella yleisin tapa kääntää suomen kielen konditionaali venäjän kieleen on käyttää ***verbiä***, jonka tapaluokkana on ***indikatiivi***, aikamuotona ***futuuri*** ja aspektina ***perfekti*.** Tämä tieto perustuu seuraaviin aineiston tilastoihin: verbillä käännettiin 887/1000 (88,7 %) esimerkkiä. Niistä 607/887 (68,4 %) esimerkkiä käännettiin indikatiivilla, joista taas 297/607 (48,9 %) esimerkkiä käännettiin futuurilla. Futuurilla käännetyistä 241/297 (81,1 %) esimerkkiä käännettiin perfektiaspektilla. Eli yhteensä tällä tavalla on käännetty 24,1 prosenttia esimerkeistä.

4.4 Aineiston analyysi käyttäen logistista regressiota

Aloitin aineistoni varsinaisen käsittelyn binäärisen logistisen regression mallinnuksella ja sen analyysillä. Kaikki käsittely tapahtui ohjelmalla `*R*` ja sen käytön helpottamista varten suunnitellulla ohjelmistoympäristöllä eli RStudiolla. Käytin R Studiossa suoritettavia toimenpiteitä varten ohjeita, jotka löytyvät apulaisprofessori Gert De Sutterin esitelmästä: Tools and Methods for Corpus-Based Translation Science. Aineiston käsittelyyn RStudiossa käytettiin seuraavia R-ohjelmiston kirjastoja: car, effects, lme4, MuMIn, MASS, party, partykit, rms, randomForest, e1071.

Tein mielenkiinnosta myös Chi-square testin aineistoni muuttujilla Kääntäjä ja Cond_RU. Testillä voidaan tarkastella seuraavatko muuttujat teoreettista jakaumaa (nollahypoteesi) vai onko olemassa joku poikkeama jakaumasta (vaihtoehtoinen hypoteesi). Tuloksena saadaan p-arvo 0,000000003183 ja tästä voi päätellä, että muuttujat seuraavat teoreettista jakaumaa, eivätkä tulokset ole sattumanvaraisia tai kääntäjien mieltymyksistä johtuvia.

Lisätietoa R ohjelmasta ja sen käytöstä voi saada esimerkiksi: R-projektin nettisivuilta osoitteesta https://www.r-project.org/ tai Natalia Levshinan (2015) teoksesta *How to do Linguistics with R. Data exploration and statistical analysis*.

4.4.1 Logistisen regressiomallin tuloksena saatavat tiedot

Selitän aluksi logistisen regressiomallin tuloksena saatavat eri tiedot ja niiden merkitykset. Logistisen regression tuloksena RStudio -ohjelmassa saadaan koostetaulukko (Kuva 5).

Ensimmäinen sarake, siis tuloksissa vasemmanpuolimmainen, pitää sisällään eri verrattavien muuttujien *vakiot* (Coefficients) eli eri arvot, joita muuttuja voi saada. Esimerkiksi siis vaikka ensimmäinen Ru_AMinf on siis Ru_AM eli venäjänkielisten käännösvastineiden aikamuodot ja yksi sen vakioista eli infinitiivi.

Kuva 5. Logistisen regression koostetaulukko (ensimmäinen malli)

```
Coefficients: (1 not defined because of singularities)
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -7.592e+01  1.769e+04  -0.004   0.9966
Ru_AMinf        1.028e-01  9.632e+03   0.000   1.0000
Ru_AMpast       8.561e-01  6.804e+03   0.000   0.9999
Ru_AMpres       1.545e+00  8.659e+03   0.000   0.9999
Ru_AMX          4.919e+01  5.465e+03   0.009   0.9928
Ru_SLadverb     7.446e-01  4.691e+04   0.000   1.0000
Ru_SLfraasi    -7.017e-01  3.470e+04   0.000   1.0000
Ru_SLgerund    -1.240e+00  3.131e+04   0.000   1.0000
Ru_SLpartik     2.573e-01  8.122e+04   0.000   1.0000
Ru_SLpartis     4.697e+01  2.647e+04   0.002   0.9986
Ru_SLpred       2.259e+01  1.683e+04   0.001   0.9989
Ru_SLsub        7.258e-01  8.122e+04   0.000   1.0000
Ru_SLverb       4.753e+01  1.720e+04   0.003   0.9978
KääntäjäIE      5.456e-01  1.331e+04   0.000   1.0000
KääntäjäMT      5.840e-01  1.630e+00   0.358   0.7201
KääntäjäOL     -2.010e+01  3.611e+03  -0.006   0.9956
KääntäjäPI     -2.070e+01  3.611e+03  -0.006   0.9954
KääntäjäSA      1.154e-01  1.055e+00   0.109   0.9129
KääntäjäTE     -9.622e-02  3.949e+04   0.000   1.0000
KääntäjäUI      1.816e+00  1.678e+00   1.083   0.2790
KääntäjäVL&IE   2.764e+00  1.680e+00   1.645   0.0999 .
KääntäjäZS&HN   8.780e-01  1.547e+00   0.568   0.5704
Ru_ASPperf      2.317e+00  1.372e+00   1.689   0.0912 .
Ru_ASPX                NA         NA      NA       NA
Fi_cndmerkmahd  3.418e+00  1.692e+00   2.020   0.0433 *
Fi_cndmerktht   9.457e-01  1.478e+00   0.640   0.5224
Fi_cndmerktvast 2.323e+00  1.636e+00   1.420   0.1556
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Toinen sarake vasemmalta on *arvio* (Estimate) vakiolle. Jos vakion arvio on negatiivinen, kuten esimerkiksi Ru_SLfraasi:n arvio -7,017e-01, niin silloin se myös vaikuttaa konditionaalin esiintymiseen negatiivisella tavalla (esiintyy vähemmän). Tässä *e*-01 tarkoittaa: kertaa 10 potenssiin -1, eli -7,017 x 10^-1, mikä taas tarkoittaa kyseisessä tapauksessa sitä, että -7,017 kerrotaan 0,1,

jolloin tulokseksi saadaan -0,7017. Luonnollisesti, jos vakion arvio on positiivinen, niin se vaikuttaa konditionaalin esiintymiseen positiivisesti (esiintyy enemmän).

Kolmannessa sarakkeessa on *keskivirhe* (Std. Error), joka nimensä mukaisesti tarkoittaa keskimääräistä virhettä. Sillä mitataan luottamusväliä, joka vastaa tiettyä todennäköisyyttä. Luottamusväli ilmaisee siis virhemarginaalin, joka sisältyy laskettuun arvioon (estimate). Eli keskivirhe kertoo virhemarginaalin. (Tilastokeskus.)

Neljännessä sarakkeessa on *standardoitu muuttujan arvo* (z value). Sillä mitataan astetta, jolla kyseinen vakio vaikuttaa konditionaalin esiintymiseen. Mitä suurempi standardoitu muuttujan arvo on, sitä vahvemmin se vaikuttaa vastemuuttujaan (konditionaalin esiintymiseen). Standardoitu muuttujan arvo saadaan, kun vakion arvio (estimate) jaetaan sen keskivirheellä (std. error). Esimerkki: ensimmäisen mallin koostetaulukossa *Fi_cndmerkmahd* saa arvion 3,418 (3,418x10^0 = 3,418) ja sen keskivirhe on 1,692 (1,692x10^0 = 1,692). Eli kun tämä lasketaan $\frac{3,418}{1,692}$ saadaan tulokseksi standardoitu muuttujan arvo 2,020. (Tilastokeskus.)

Viimeisenä oikeanpuolimmaisessa sarakkeessa on vakion *p-arvo* (Pr (>|z|)) eli havaittu merkitsevyystaso. Mitä alempi p-arvo on, sen varmempi voi olla siitä, että nollahypoteesi on väärässä ja vaihtoehtoinen hypoteesi on oikeassa eli näytteessä on poikkeama jakaumasta kyseisen vakion osalta. P-arvon raja on yleensä 5 prosenttia ($p \leq 0.05$). Toisin sanottuna tämä tarkoittaa sitä, että nollahypoteesin todennäköisyys on ainoastaan 5 % ja vaihtoehtoisen hypoteesin todennäköisyys on 95 %. Koostetaulukoissa alimmalla rivillä on selitykset p-arvojen vieressä näkyville merkinnöille (Signif. codes:), esimerkiksi `*` p-arvon vieressä tarkoittaa, että p-arvo alittaa 5 prosentin rajan. Tämä auttaa hahmottamaan helposti, mitkä regressiomallin muuttujien vakiot vaikuttavat vastemuuttujaan

4.4.2 Aineiston binäärinen regressiomallinnus

Tavoitteenani on siis selvittää, mitkä tekijät vaikuttavat siihen, että venäjänkielinen konditionaalissa. Toisin sanottuna vastemuuttujana minulla on binäärinen muuttuja *Cond_RU* eli onko käännösvastine konditionaali vai ei (*Yes* tai *No*). Selittäviksi muuttujiksi otan aluksi lähes kaikki muut aineistoni muuttujat: *Kääntäjä*, suomenkielinen konditionaalimuodon verbi *Fi_form*, konditionaaliesimerkin tapaluokka ja aikamuoto *Fi_TL_AM*, konditionaalimerkitys *Fi_cndmerk*, verbin semanttinen merkitys *Fi_vclas*, käännösvastine *Ru_form*, käännösvastineen sanaluokka *Ru_SL*, käännösvastineen aikamuoto *Ru_AM*, käännösvastineen aspekti *Ru_ASP* ja Leks_ru. Jätän siis pois muuttujat *Kirjailija* ja *Teos_fi* eli teoksen nimi, sillä nämä eivät voi mitenkään vaikuttaa käännösvastineen tapaluokkaan. Myös muuttuja *Ru_TL* eli käännösvastineen tapaluokka jäi luonnollisesti pois tästä selittävien

muuttujien listasta, koska vastemuuttujani on sen pohjalta luotu. Otin ensimmäiseen regressiomalliin mukaan selittäviksi muuttujiksi näin suuren määrän muuttujia, koska en voi olla varma, mikä näistä vaikuttaa konditionaalin esiintymiseen käännöksissä, ja halusin tarkistaa asian. Itse komento jota käytin tässä ensimmäisessä regressiomallinnuksessa oli *MASS* -kirjaston funktio *stepAIC(glm())*.

Ensimmäisen regressiomallin tuloksessa (Kuva 5) ainoastaan suomenkielisen esimerkin konditionaalimerkitys *mahd*, eli intensionaalista asiantilaa ilmaiseva merkitys, alittaa 5 prosentin p-arvon eli saa merkittävän tuloksen eli 0.0433. Ensimmäisen mallin tuloksessa vakio *Fi_cndmerkmahd* saa arvion (estimate) 3,418, joten se vaikuttaa positiivisella tavalla konditionaalin esiintymiseen käännöksissä ja sen standardoitu muuttujan vakio (z value) on myös suuri 2,020 eli se vaikuttaa siihen suuresti. Muuttujista *Fi_form*, *Ru_form* ja *Leks_ru* eivät edes näy koostetaulukossa eli voin olla varma, että yksikään niiden vakioista ei myöskään vaikuta konditionaalin esiintymiseen. Muuttuja *Ru_SL* ja sen eri vakiot esiintyvät koostetaulukossa, mutta yksikään vakioista ei saa merkittävää p-arvoa. Muuttuja *Ru_AM* ja sen vakiot esiintyvät myös koostetaulukossa, mutta eivät saa merkittäviä p-arvoja. Uskon kuitenkin, että venäjänkielisen käännösmuodon aikamuoto voi jollain tapaa vaikuttaa konditionaalin esiintymiseen, joten jätän sen vielä seuraavaan malliin. Loput kolme regressiomallia tein *MASS* -kirjaston komennolla *glm()*.

Näiden seikkojen perusteella muokkaan regressiomallia poistamalla siitä seuraavat selittävät muuttujat: *Fi_form, Ru_form, Ru_SL ja Leks_ru*. Muokatussa, toisessa regressiomallissa, on siis selittävinä muuttujina jäljellä: *Kääntäjä, Fi_TL_AM, Fi_cndmerk, Fi_vclas, Ru_AM* ja *Ru_ASP*.

Toisen regressiomallin tuloksessa merkittäviä p-arvoja saivat jo kolme vakiota eli *KääntäjäPI* eli Prileżaev, Ivan (0.00293), *Fi_vclasSUH* eli asioiden suhteita ilmaisevat verbit (0.02949) ja *Ru_ASPperf* eli käännösvastineen perfektiaspekti (0.00609).

Muokattu malli antaa siis jo paremman tuloksen kuin ensimmäinen, mutta muokkaan sitä lisää poistamalla siitä vielä selittävän muuttujan *Ru_AM*, sillä yksikään *Ru_AM* vakioista ei edelleenkään saa merkittävää p-arvoa. Kolmas malli koostuu siis seuraavista selittävistä muuttujista: Kääntäjä, *Fi_TL_AM*, *Fi_cndmerk*, *Fi_vclas* ja *Ru_ASP*. Jätän tähän kolmanteen malliin muuttujan Fi_TL_AM, koska uskon sen ainakin varmasti vaikuttavat konditionaalin esiintymiseen, vaikka sen vakiot eivät saa aiemmissa malleissa merkittäviä p-arvoja.

Kolmannen regressiomallin tuloksessa muuttujat *Kääntäjä, Fi_TL_AM ja Fi_cndmerk,* vaikuttavat vahvasti konditionaalin esiintymiseen ja niiden eri vakiot saavat merkittäviä p-arvoja. *Ru_ASPX* (eli aspektiton) on ainut *Ru_ASP* muuttujan vakioista, joka saa merkittävän tuloksen, joten *Ru_ASP* poistetaan seuraavasta mallista. Myös verbien semanttisen merkityksen muuttujan *Fi_vclas* vakioista

ainoastaan *Fi_vclasLII* (eli liikettä ilmaisevat verbit) saa merkittävän tuloksen, joten myös *Fi_vclas* poistetaan seuraavasta regressiomallista.

Neljäs, ja mielestäni optimaalinen, regressiomalli sisältää siis jäljelle jääneet selittävät muuttujat: *Kääntäjä, Fi_TL_AM ja Fi_cndmerk*. Tästä saadussa tuloksessa kaikki muuttujat vaikuttavat, jollain tavalla *Cond_RU* vastemuuttujaan (kuva 6). Kääntäjä -muuttujan 3 eri vakiota saavat merkittävän p-arvon. Kaksijakoisen *Fi_TL_AM* -muuttujan toinen vakio *cnd_pres* saa erittäin merkittävän p-arvon (0,000000000814). *Fi_cndmerk* -muuttujan 2 eri vakiota saavat merkittävät p-arvot.

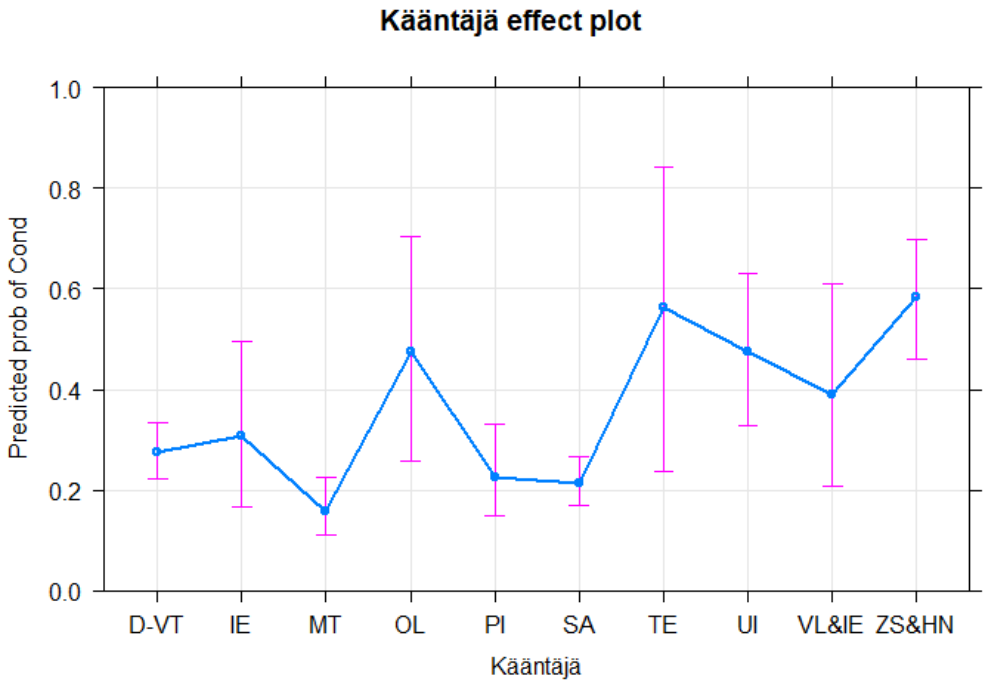
Kuva 6. Neljännen regressiomallin koostetaulukko

```
Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)       0.8499     0.2508   3.388 0.000703 ***
KääntäjäIE        0.1521     0.4306   0.353 0.723907
KääntäjäMT       -0.6979     0.2603  -2.681 0.007343 **
KääntäjäOL        0.8712     0.5090   1.711 0.087001 .
KääntäjäPI       -0.2589     0.3029  -0.855 0.392644
KääntäjäSA       -0.3318     0.2048  -1.621 0.105088
KääntäjäTE        1.2188     0.7348   1.659 0.097165 .
KääntäjäUI        0.8720     0.3504   2.489 0.012817 *
KääntäjäVL&IE     0.5200     0.4776   1.089 0.276239
KääntäjäZS&HN     1.3042     0.2914   4.476 7.59e-06 ***
Fi_TL_AMcnd_pres -1.0772     0.1754  -6.142 8.14e-10 ***
Fi_cndmerkmahd   -1.2029     0.2477  -4.856 1.20e-06 ***
Fi_cndmerktht    -0.6289     0.3734  -1.684 0.092115 .
Fi_cndmerktvast  -1.1842     0.2003  -5.911 3.40e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

4.4.3 Kääntäjän vaikutus konditionaalin käyttöön käännöksissä

R-ohjelmalla on mahdollista piirtää regressiomallin tulokset diagrammeiksi, mikä helpottaa tulosten tarkastelua. Tämä tehdään komennolla *plot(allEffects())* ja tulokseksi saadaan seuraavanlaisia kuvioita (kuvat 7-9).

Kuva 7. Konditionaalin käytön ennustettu todennäköisyys *Kääntäjä*



Kääntäjä -muuttujan ennustettu konditionaalin käytön todennäköisyys on merkattuna jokaisen kääntäjän kohdalle erikseen (kuva 7). Alhaalla (X-akseli) on siis lyhenteet kääntäjien nimistä ja vasemmalla sivulla (Y-akseli) on ennustettu konditionaalin käytön todennäköisyys (Predicted probability of Cond). Violetit pystyviivat näyttävät virhemarginaalin (Std. Error) suuruuden suhteessa ennustettuun konditionaalin käytön todennäköisyyteen.

Diagrammissa voidaan nähdä melko suuria eroja kääntäjien konditionaalimuodon käytön ennusteissa. Alla olevassa taulukossa (Taulukko 11) on merkittynä jokaisen kääntäjän kääntämien esimerkkien määrä, joka on jaettu konditionaalilla käännettyihin (Yes) ja millä tahansa muulla käännösvastineella käännettyihin (No).

Logistinen regressio ja sen tuloksena saamani diagrammi (kuva 7) on mielestäni hyvä tapa visualisoida ja analysoida aineistoni dataa. Esimerkiksi kääntäjän **TE** (Tinovickaâ, Evgeniâ) ennustettu konditionaalin käytön todennäköisyys on hyvin korkea, mutta diagrammissa näkyy myös, että virhemarginaali on valtava (0.7348). Tämä johtuu luonnollisesti pienestä esimerkkien

käännösmäärästä (8 kappaletta). Toisin sanottuna kääntäjän **TE** osalta on aineistoni perusteella oikeastaan mahdotonta tehdä päätelmiä hänen konditionaalin käytöstään käännösratkaisuna. Otetaan tarkasteluun mieluummin kääntäjä, joka on kääntänyt suurimman aineistoni osan konditionaaliesimerkeistä: **SA** (Sidorova, Anna). Kuten diagrammista (kuva 7) ja koostetaulukosta (kuva 7) nähdään, on SA:n virhemarginaali aineistoni pienin: 0.2048. Hänen konditionaalin käyttönsä ennuste on kuitenkin hyvin pieni, koska aineistossa hän on kääntänyt vain 22,9 prosenttia konditionaalimuodoista konditionaalilla (Taulukko 11).

Sama asia voidaan nähdä tietenkin myös neljännen regressiomallin koostetaulukosta (kuva 7), jossa *KääntäjäSA* saa arvion -0.3318. **SA**:n osalta on mahdollista tehdä päätelmiä hänen konditionaalin käytöstä käännöksissä, koska hänen kääntämien esimerkkien määrä on mielestäni riittävä tätä varten (297 kappaletta). En kuitenkaan aio tutkia erillisten kääntäjien konditionaalin käyttöä sen tarkemmin, vaan tarkastelen sitä yleisemmällä tasolla.

Taulukko 11. Kääntäjien konditionaalin käyttö käännöksissä.

Kääntäjä	Yes	No	Esimerkkimäärä
Sidorova, Anna (**SA**)	**68** (22,90%)	**229** (77,10%)	**297**
Džafarova-Viitala, Tais'â (**D-VT**)	**74** (27,92%)	**191**(72,08%)	**265**
Mel′nik, Tat′âna (**MT**)	**28** (17,28%)	**134** (82,72%)	**162**
Prileżaev, Ivan (**PI**)	**24** (29,63%)	**57** (70,37%)	**81**
Zajkov, S. and Hotinskij, N. (**ZS&HN**)	**39** (56,52%)	**30** (43,48%)	**69**
Ureckij, Il′â (**UI**)	**21** (47,73%)	**23** (52,27%)	**44**
Ioffe, Èleonora (**IE**)	**9** (27,27%)	**24** (72,73%)	**33**
Virolajnen, Laura and Ioffe, Èleonora (**VL&IE**)	**9** (40,91%)	**13** (59,09%)	**22**
Olykajnen, Leonid (**OL**)	**10** (52,63%)	**9** (47,37%)	**19**
Tinovickaâ, Evgeniâ (**TE**)	**4** (50%)	**4** (50%)	**8**
Yhteensä (Keskiarvo)	**286** (37,27%)	**714** (62,72%)	**1000**

Taulukossa jako on kuvattu myös suluissa prosenteissa ja alimmalla rivillä on suluissa prosenttimäärillä kuvattuna kaikkien kääntäjien konditionaalin käytön keskiarvo. Jos tämän perusteella haluaisi tehdä johtopäätöksiä konditionaalin käytöstä, niin voisi sanoa, että konditionaalilla käännetään venäjään keskimäärin 37 prosenttia suomenkielisistä konditionaalimuodoista. Tämä ei kuitenkaan ole totta, koska jos tarkastellaan taulukkoa niin huomataan, että mitä enemmän käännöksiä kääntäjä on tehnyt, sen pienemmäksi konditionaalilla tehtyjen käännösten osuus keskimäärin muuttuu. Katsotaan esimerkiksi neljän eniten käännöksiä aineistossani tehneen kääntäjän konditionaalin käytön jakaumaa eli kääntäjät **SA, D-VT, MT** ja **PI**, jotka ovat yhteensä kääntäneet 805 aineistoni tuhannesta konditionaaliesimerkistä. Näiden neljän

kääntäjän konditionaalin keskimääräinen käyttö suomenkielisen konditionaalimuodon kääntämisessä on ainoastaan 24,43 prosenttia eli 75,57 prosenttia on käännetty jollain muulla tavalla.

Tämän kaavan rikkoo kuitenkin yksi kääntäjä(pari): **ZS&HN** (Zajkov, S. & Hotinskij, N.). He ovat kääntäneet aineistoni esimerkeistä yhteensä 69 kappaletta. Näistä 39 (56,52 %) on käännetty konditionaalilla ja 30 jollain muulla tavalla. Esimerkki aineistostani ZS&HN konditionaalin käytöstä käännöksessä (esimerkki 13).

(13) Helppo tämmöistä vanhaa akkaa oli metsissä riepotella! Jos Naska **_olisi_** nuorempi, niin ei häntä tällä lailla retuutettaisi. (A. Paasilinna, Hirtettyjen kettujen metsä) / Нетрудно такую старуху по лесам таскать! Если **бы** Наска **была** помоложе, то ее так не волочили бы. (Les povešennyh lisic, kääntäjät S. Zajkov & N. Hotinskij)

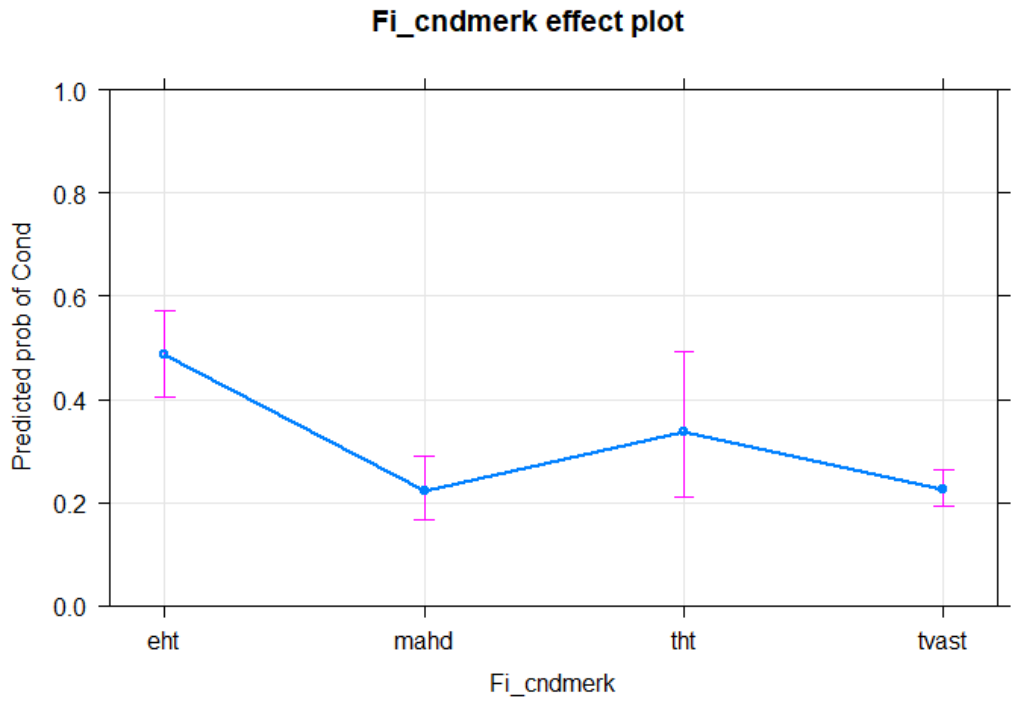
Diagrammin (kuva 7) mukaan konditionaalin käytön todennäköisyyden ennuste on heidän osaltaan hyvin korkea, lähes 0,6 (60 %). Mielenkiintoista, että binäärisen logistisen regression tuloksessa ennuste on vielä suurempi, kuin 56,52 %, mikä oli heidän aito konditionaalin käyttönsä. Virhemarginaalikaan ei ole kovin suuri vain (0.2914), koska esimerkkejä on kuitenkin 69 kappaletta. Vaikea kuitenkaan sanoa johtuuko tämä tulos siitä, että ZS&HN kääntävät suomen kielen konditionaalin useammin konditionaalilla venäjään kuin muut, vai onko kyse vain tilastollisesta poikkeamasta, joka aineiston kasvaessa esimerkiksi 10 000 konditionaaliesimerkkiin tasaantuisi ja alkaisi vastata tuota 25/75 % jakaumaa (25% Yes, 75% No), joka on neljällä eniten käännöksiä tehneellä kääntäjällä.

Näiden binäärisen logistisen regression tulosten perusteella voin tehdä päätelmiä Kääntäjä -muuttujan vaikutuksesta konditionaalin käyttöön käännöksissä. Kääntäjä henkilönä vaikuttaa tietenkin konditionaalin esiintymiseen käännöksissä, koska hän tekee käännösratkaisuja eri kielimuotojen kääntämiseen ja saattaa suosia jotain tiettyä käännösratkaisua. Aineistoni perusteella voisi sanoa esimerkiksi kääntäjän ZS&HN suosivan suomen kielen konditionaalimuodon kääntämiseen käännösratkaisuna venäjän kielen konditionaalimuotoa. Teoretisoin kuitenkin, että aineiston esimerkkimäärän kasvaessa konditionaalin käyttö tasaantuisi kääntäjästä riippumatta, ja päätyisi vastaamaan jakaumaa, jossa 25 % käännetään konditionaalilla ja 75 % muilla käännösratkaisuilla. Tätä tukee myös aineistoni käännösten konditionaalin käytön jakauma, koska kuten aiemmin kerroin aineistoni esimerkeistä käännettiin 28,6 % konditionaalilla.

4.4.4 Fi_cndmerk vaikutus konditionaalin käyttöön käännöksissä

Seuraavana diagrammi ennustetusta konditionaalin käytön todennäköisyydestä muuttujan Fi_cndmerk vakioiden yhteydessä (kuva 8). Eri vakioita ovat luonnollisesti ne 4 eri konditionaalimerkitystä, joihin suomen kielen konditionaali jakautuu ja joilla jaoin oman aineistoni (eht, mahd, tht, tvast). Ei ole yllättävää, että suurin virhemarginaali on näistä vakiolla tht (tahtotulkintainen), jota oli reilusti vähiten aineistossani, vain 48 kpl.

Kuva 8. Konditionaalin käytön ennustettu todennäköisyys *Fi_cndmerk*



Kuten mainitsin tässä tutkielmassa (luku 3.3), sekä suomen että venäjän konditionaalilla voidaan ilmaista ehtoa. Venäjän kielessä konditionaali ilmaisee asiantilaa, joka ei ole olemassa reaalimaailmassa. Venäjässä asiantila ei siis voi olla eikä tulla koskaan olemaan oikeassa maailmassa ja suomen kielessä konditionaalilla kuvailtava asiantila on yksi mahdollisista vaihtoehdoista. On siis loogista, että diagrammi *Fi_cndmerk* (kuva 8), näyttää, miten ehtotulkintaisen konditionaalimerkityksen (eht) omaavat konditionaaliesimerkit on käännetty huomattavasti useammin konditionaalilla venäjään, kuin esimerkiksi intensionaalista asiaintilaa ilmaisevan konditionaalimerkityksen (mahd) omaavat konditionaaliesimerkit. Uskoakseni juuri tästä syystä aineistossani intensionaalista asiantilaa (mahd) kuvaavien konditionaalimerkitysten osalta 193:sta esimerkistä ainoastaan 46 kappaletta (23,8 %) kääntyykin konditionaalilla (Taulukko 12). Ei ole siis ihmeellistä, että ennustettu todennäköisyys konditionaalin käytölle on korkea ainoastaan ehtotulkintaisella konditionaalimerkityksellä. En huomioi tässä tahtotulkintaisen

konditionaalimerkityksen ennustettua konditionaalin käytön todennäköisyyttä, sillä sen omaavien esimerkkien määrä on kovin pieni (48 kpl), ja koska sen jakauma on pienestä määrästä riippumatta silti lähes *tvast* tai *mahd* vastaava (eli 31 % Yes, 68 % No). Sen virhemarginaali on myös melko suuri eli 0,3734.

Taulukko 12. Konditionaalijakauma eri konditionaalimerkityksissä

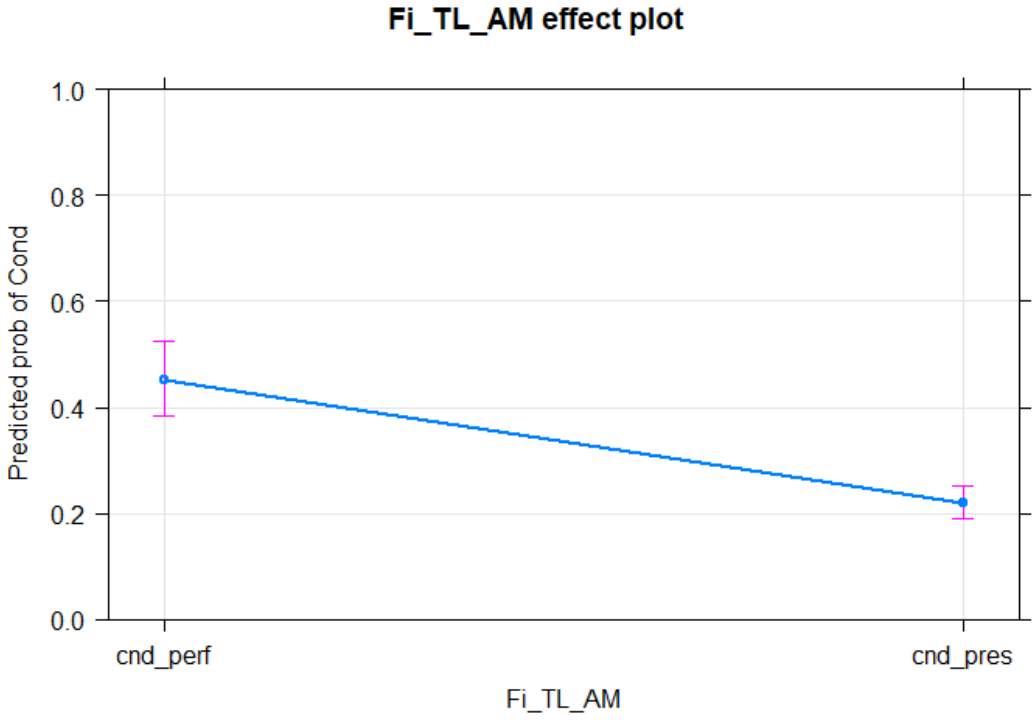
Fi_cndmerk – Onko konditionaali?	**Lukumäärä**	**%**
eht	**153**	**15,3**
No	73	47,7
Yes	80	52,3
mahd	**193**	**19,3**
No	147	76,2
Yes	46	23,8
tht	**48**	**4,8**
No	33	68,8
Yes	15	31,2
tvast	**606**	**60,6**
No	461	76
Yes	145	24
Yhteensä	**1000**	**100**

Loppupäätelmänä voisi siis sanoa, että optimaaliseksi valitsemani regressiomallin tulosten koostetaulukossa kaikki muut kolme konditionaalimerkitystä (mahd, tht, tvast) saavat negatiivisen *arvion* (estimate) ja *standardisoidun muuttujan arvon* (z value) eli ne vaikuttavat konditionaalin esiintymiseen negatiivisesti. Toisin sanottuna aineistoni perusteella voisin väittää, että konditionaalimerkityksellä on siis vaikutusta siihen, kääntyykö suomen kielen konditionaali venäjään konditionaalilla, mutta ainoastaan ehtotulkintaisen (eht) konditionaalimerkityksen omaavat suomen kielen konditionaalimuodot kääntyvät useammin venäjän kielen konditionaalilla, kuin muilla käännösratkaisuilla. Muiden osalta vaikutus on siis päinvastainen eli kääntyvät harvemmin venäjän kielen konditionaalilla.

4.4.5 Fi_TL_AM vaikutus konditionaalin käyttöön käännöksissä

Neljännen binäärisen logistisen regressiomallin diagrammi *Fi_TL_AM* eli suomenkielisten esimerkkien tapaluokkaa ja aikamuotoa kuvastava muuttuja (kuva 9). Tässä näkyy ennustettu konditionaalin käytön todennäköisyys eri *aikamuotojen* (AM) yhteydessä, koska kaikki esimerkit ovat tietenkin konditionaalissa.

Kuva 9. Konditionaalin käytön ennustettu todennäköisyys *Fi_TL_AM*



Venäjän kielen konditionaalimuodoista huomattavasti yleisimpiä ovat verbin perfektimuodon (***-l-muoto***, kts. luku 3.2) ja partikkelin *by (b)* yhdistelmät. Ennustettu konditionaalin käytön todennäköisyys on huomattavasti korkeampi silloin, kun suomenkielinen konditionaaliesimerkki on perfektissä (*cnd_perf)*. Tätä tukevat myös koostetaulukossa (kuva 7) nähtävät tiedot, joissa *Fi_TL_AMcnd_pre*s -vakio saa negatiivisen arvion (-1.0772) ja erittäin ison standardisoidun muuttujan arvon (-6.142), eli se vaikuttaa vahvasti negatiivisella tavalla konditionaalin esiintymiseen käännöksissä.

Logistisen regression tuloksen diagrammista (kuva 9) voidaan siis nähdä selvä ero konditionaalin käytön ennustetussa todennäköisyydessä näiden kahden suomenkielisten esimerkkien aikamuodon välillä. Konditionaalin perfektimuotoa (223) on suomenkielisissä esimerkeissä huomattavasti vähemmän kuin konditionaalin preesensmuotoa (777).

En valitettavasti pysty tässä tarkistamaan, kuinka suuri osa aineistoni konditionaalilla käännetyistä venäjänkielisistä käännösvastineista koostuu juuri perfektimuotoisen verbin ja *by* partikkelin yhdistelmistä. Tämä johtuu siitä, että venäjän kielessä konditionaalilla ei ole aikamuotoa, koska se ei erota aikamääreitä vaan tilanne, jota konditionaali kuvastaa voi olla samanaikaisesti preesens, perfekti ja futuuri. Esimerkkejä käsitellessä olen siis konditionaalin esiintyessä käännöksessä laittanut Ru_AM sarakkeeseen vastaavalle riville X (ei saatavilla). Näitä oli 359 kappaletta, kuten alla olevasta taulukosta (taulukko 13) voidaan nähdä.

Konditionaalin perfektin 114:sta aikamuodottomasta suomenkielisestä esimerkistä 99 kappaletta (86,8 %) on venäjänkielisessä käännöksessä konditionaalimuodossa. Konditionaalin preesensin 245:stä aikamuodottomasta suomenkielisestä esimerkistä 187 kappaletta (76,3 %) on venäjänkielisessä käännöksessä konditionaalissa. Tällä tavalla vertailtaessa ero on ainoastaan noin 10 prosenttia. Ero kuitenkin kasvaa huomattavasti, jos otetaan huomioon se, kuinka monta suomenkielistä esimerkkiä on yhteensä kummassakin aikamuodossa (eli cnd_pres ja cnd_perf). Tässä tapauksessa konditionaalin perfektissä olleet suomenkieliset konditionaaliesimerkit kääntyvät 44 prosenttia (99/223) ajasta konditionaalilla myös venäjän kieleen (taulukossa 13: *Yes % Fi_TL_AM)*. Toisin kuin konditionaalin preesensissä olleet konditionaaliesimerkit, jotka kääntyvät konditionaalilla venäjään vain 24:ssä (187/777) prosentissa tapauksista.

Taulukko 13. Fi_TL_AM jakauma venäjänkielisten käännösten aikamuotoihin

Fi_TL_AM	Ru_AMpres	Ru_AMpast	Ru_AMfut	Ru_AMinf	Ru_AMX	Cond_RU Yes	Yes % Fi_TL_AM
cnd_perf	9	79	7	14	114	99	44 %
cnd_pres	81	97	296	58	245	187	24 %
Yhteensä	**90**	**176**	**303**	**72**	**359**	**286**	

Voin näiden tietojen perusteella väittää, että suomenkielisen konditionaalimuodon aikamuoto vaikuttaa vahvasti siihen, että kääntyykö se venäjän kieleen konditionaalilla vai ei. Esimerkin ollessa perfektiaikamuotoinen, on aineistoni perusteella lähes 2 kertaa todennäköisempää, että myös käännös tehdään konditionaalilla.

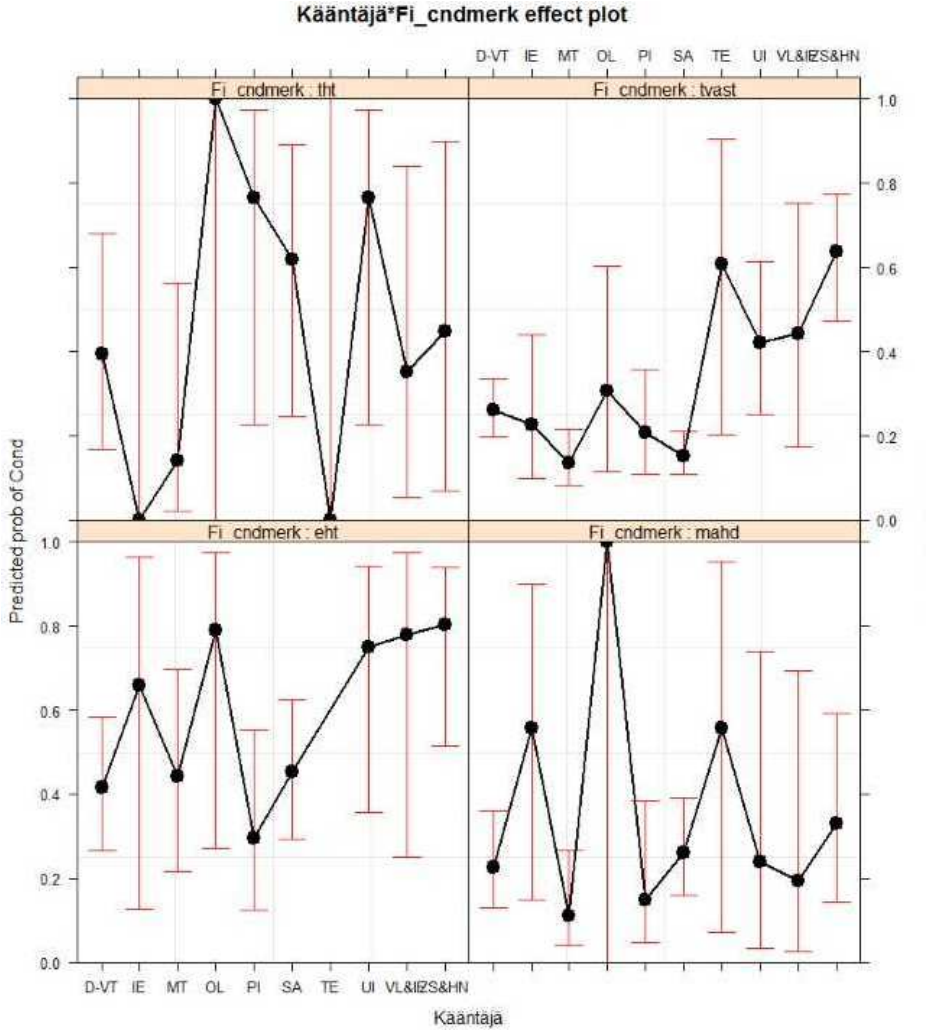
4.4.6 Muuttujien *Kääntäjä, Fi_cndmerk* ja *Fi_TL_AM* vaikutus toisiinsa

Tarkistetaan vaikuttavatko selittävät muuttujat toisiinsa. Diagrammista (Kuva 7) ja taulukosta (taulukko 11) voidaan nähdä, kuinka jotkut kääntäjistä saattavat usein käyttää tietyn kielimuodon kääntämistä varten samaa käännösratkaisua enemmän kuin toiset, esim. **ZS&HN**, jotka käyttivät suomen kielen konditionaalin kääntämiseen venäjän kielen konditionaalia 56 prosentissa käännöksistään. Mielenkiintoista on se, että aineistossani ei esiinny ainuttakaan esimerkkiä, jossa Fi_TL_AM on *cnd_perf* ja Fi_cndmerk on *tht*. Tahtotulkintaisen konditionaalin perfektimuoto ilmaisee vaihtoehtoista tapahtumien kulkua jo menneille tapahtumille (kts. luku 3.1). Saattaa siis olla, että tämä ei ole kovinkaan yleinen konditionaalimerkityksen ja aikamuodon yhdistelmä kaunokirjallisuudessa tai se ei vain satu esiintymään aineistossani.

Tarkastellaan siis binäärisen logistisen regression tuloksena saatavia diagrammeja (kuva 10) selittävien muuttujien *Kääntäjä* ja *Fi_cndmerk* vaikutuksesta toisiinsa ja konditionaalin käytön todennäköisyyden ennustetta, jonka ne saavat. Kääntäjistä **TE** on ainut, kuka ei ole kääntänyt yhtäkään konditionaaliesimerkkiä, jossa konditionaalimerkitys olisi ollut ehtotulkintainen (eht). Siitä, johtuu ennusteen puuttuminen hänen kohdallaan *Fi_cndmerk: eht* -diagrammissa. Kaikki muut kääntäjät ovat kääntäneet vähintään yhden esimerkin kaikkia eri konditionaalimerkityksiä ilmaisevia konditionaalimuotoja.

Jätän tässä analyysissä huomioimatta tahtotulkintaisen (tht) konditionaalimerkityksen ja kääntäjien saamat ennusteet, koska aineistoni konditionaaliesimerkeistä ainoastaan 4,8 prosenttia on tätä konditionaalimerkitystä edustavia. Ennusteet heittelevät pienen määrän takia ja täten on mielestäni aivan turha tehdä mitään johtopäätöksiä siitä, onko tahtotulkintaisen konditionaalin esiintymisellä vaikutusta siihen kääntääkö kääntäjä konditionaalilla vai ei. Mainittakoon kuitenkin, että kaikista konditionaaliesimerkeistä, joiden konditionaalimerkitys on *tht,* kääntäjät käänsivät 31,25 % konditionaalilla. Kuvien 10 ja 11 diagrammit piirretään samalla R-ohjelman komennolla kuin aiemmatkin kuviot eli *plot(allEffects().*

Kuva 10. Selittävien muuttujien Kääntäjä ja Fi_cndmerk vaikutus toisiinsa ja ennusteet konditionaalin käytön todennäköisyydelle



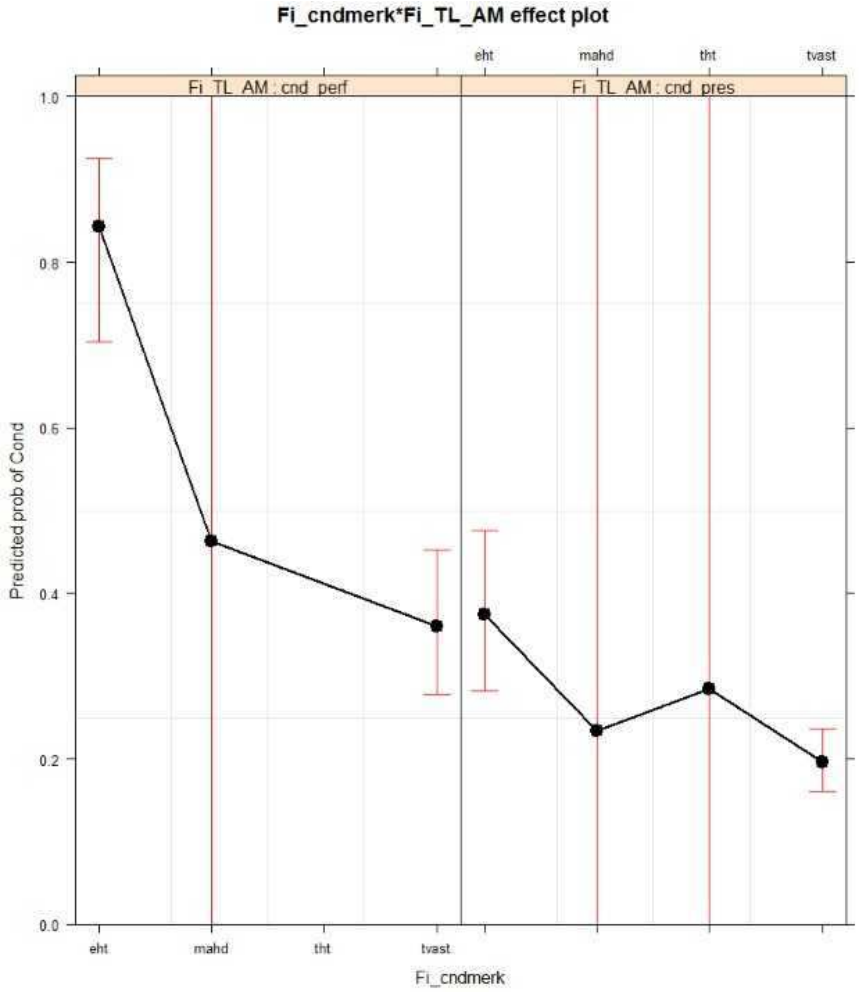
Eri kääntäjillä on luonnollisesti omat preferenssit konditionaalin käytön suhteen. Diagrammeista (kuva 10) päätellen konditionaalimerkityksellä on kuitenkin vaikutusta siihen, käyttääkö kääntäjä käännösratkaisuna venäjän kielen konditionaalia vai ei. Konditionaaliesimerkit, joissa konditionaalimerkitys on intensionaalista asiaintilaa ilmaiseva (mahd), saavat keskimäärin huomattavasti alemman kääntäjien konditionaalin käytön todennäköisyyden ennusteen kuin ehtotulkintaiset konditionaaliesimerkit. Konditionaaliesimerkit, joissa konditionaalimerkitys on todenvastainen (tvast) saavat keskimäärin myös alhaisen kääntäjien konditionaalin käytön todennäköisyyden ennusteen ja niiden lukumäärä myös on huomattavasti suurempi (60 % aineiston konditionaalimerkityksistä). Kummankaan konditionaalimerkityksen osalta tämä ei ole yllättävää, koska näistä molemmista konditionaalimerkityksistä kääntäjät käänsivät alle neljäsosan (mahd = 23,8 %, tvast = 24 %) konditionaalilla. Näihin verrattuna kääntäjät käyttivät ehtotulkintaisten konditionaalimerkitysten käännöksissä 52 prosentissa käännösratkaisuna konditionaalia.

Diagrammien tulosten seassa on muutama tilastollinen poikkeama, jotka vääristävät niitä. Esimerkiksi *Fi_cndmerk : mahd* -diagrammissa kääntäjä **OL** (Olykajnen, Leonid) saa konditionaalin käytön todennäköisyyden ennusteeksi 1.0 (100 %). Tämä johtuu siitä, että kyseinen kääntäjä on kääntänyt ainoastaan yhden konditionaaliesimerkin, jonka konditionaalimerkitys on *mahd*, ja hän käänsi sen venäjän kielen konditionaalilla. Kääntäjä **ZH&HN** saa tämänkin mallin myötä, konditionaalimerkityksestä riippumatta, korkean ennusteen konditionaalin käytön todennäköisyydelle, koska tämä on ainoa kääntäjä(pari), kuka on kääntänyt enemmän konditionaalilla kuin muilla tavoilla (56,52 % käännöksistä).

Diagrammeissa (kuva 10) voi nähdä, että tulosten virhemarginaalit ovat suurimmaksi osin melko valtavia. Tämän voisi korjata, jos kasvattaisi aineiston esimerkkimäärää muutamalla tuhannella tai muokkaamalla aineiston hakua niin, että hakutuloksessa eivät olisi hieman yliedustettuina tietyt kääntäjät. On esimerkiksi vaikea tutkia eri kääntäjien mieltymyksiä tai käännösratkaisuja ja vertailla niitä, kun yksi kääntäjä **IE** (Ioffe, Èleonora) on kääntänyt aineiston esimerkeistä 33 (3,3 %) ja toinen **D-VT** (Džafarova-Viitala, Tais′â) on kääntänyt 265 (26,5 %) esimerkkiä.

Kuitenkin näiden logistisen regressiomallin tulosten perusteella vahvistuu jo aiemmin mainitsemani asia: jos konditionaaliesimerkki on ehtotulkintaista konditionaalia ilmaiseva, niin on todennäköisempää, että se käännetään venäjän kielen konditionaalilla (52 % ajasta). Tutkittaessa näiden selittävien muuttujien vaikutusta toisiinsa, voidaan myös väittää, että tämä ei ole kääntäjästä riippuva asia, koska kaikkien muiden kääntäjien ennuste konditionaalin käytön todennäköisyydelle ehtotulkintaista konditionaalia käännettäessä on yli 40 prosenttia eli kyseessä on yleinen ilmiö. Ehtotulkintaisen konditionaalin *Fi_cndmerk : eht* diagrammissa alle 40 % todennäköisyyden konditionaalin käytölle saa ainoastaan kääntäjä PI. Kyseinen kääntäjä on kääntänyt 20 ehtotulkintaista konditionaaliesimerkkiä, joista 10 konditionaalilla ja 10 muilla tavoin.

Kuva 11. Selittävien muuttujien Fi_cndmerk ja Fi_TL_AM vaikutus toisiinsa ja ennusteet konditionaalin käytöstä



Kuten tämän luvun (4.4.6) alussa mainitsin aineistossani ei esiinny ainuttakaan esimerkkiä, jossa Fi_TL_AM on *cnd_perf* ja Fi_cndmerk on *tht*. Tästä johtuu tietenkin se, että kyseisen konditionaalimerkityksen kohdalla ei ole ennustetta konditionaalin käytön todennäköisyydelle diagrammissa *Fi_TL_AM : cnd_perf* (kuva 11).

Tarkastellessa diagrammeja selittävien muuttujien *Fi*_cndmerk ja *Fi_TL*_AM vaikutuksesta toisiinsa, huomaa kuitenkin, että ne näyttävät saman asian, joka näkyy logistisen regressiomallin aiemmissa tuloksissa. Ehtotulkintainen konditionaalimerkitys saa molemmissa aikamuodoissa korkeimman konditionaalin käytön todennäköisyyden ennusteen. Todenvastainen konditionaalimerkitys (tvast) saa alimman ennusteen molemmissa aikamuodoissa ja intensionaalista asiantilaa ilmaiseva konditionaalimerkitys (mahd) saa molemmissa aikamuodoissa ennusteen näiden kahden väliltä.

Tästä diagrammista nähdään myös, että aikamuodolla on, konditionaalimerkityksestä riippumatta, vaikutus konditionaalin käyttöön käännösratkaisuna. Eli kaikki konditionaalimerkitykset saavat

korkeamman konditionaalin käytön todennäköisyyden ennusteen, kun konditionaaliesimerkin aikamuoto on perfekti.

Tämä diagrammi selittävien muuttujien vaikutuksesta toisiinsa todistaa selkeällä tavalla aiemmat huomioni: Ehtotulkintainen konditionaali kääntyy useimmin konditionaalilla (52 % ajasta) ja perfektimuotoinen konditionaali kääntyy myös usein konditionaalilla (44 % ajasta). Tämä diagrammi näyttää siis saman asian yhdistettynä ja antaakin ennusteeksi yli 80 % todennäköisyyden konditionaalin käytölle, silloin kun nämä molemmat esiintyvät samassa konditionaaliesimerkissä.

4.5 Ehdollisen päättelyn puu ja Satunnainen metsä

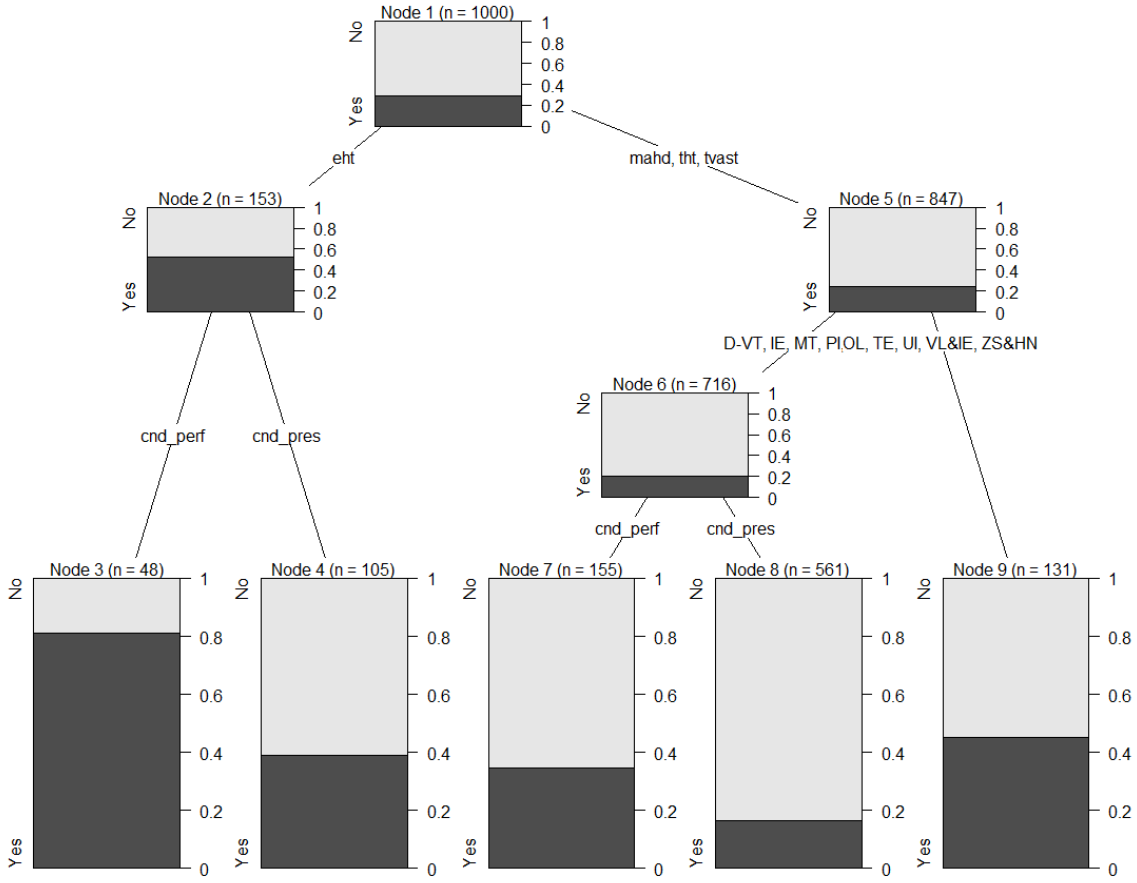
4.5.1 Ehdollisen päättelyn puun analysointi

Tein optimaaliseksi valitsemieni logistisessa regressiossa käyttämieni selittävien muuttujien perusteella aineistostani ehdollisen päättelyn puun. Käytin tähän R:ssä *ctree()* funktiota, joka löytyy R:n kirjastosta *party*. Vastemuuttujana minulla oli tässä tietenkin *Cond_RU* ja selittävinä muuttujina *Kääntäjä, Fi_TL_AM ja Fi_cndmerk*.

Kuvassa 12 on kuvattuna siis ehdollisen päättelyn puu, jossa näkyy kaikki ositukset, jotka ovat tärkeysasteeltaan 0.05 alittavia. Puun pylväskuvioissa (noodeissa) näkyvät ositukseen valittujen selittävien muuttujien jakaumat jaettuna vastemuuttujalla ja niiden vastaavat p-arvot, jotka ovat kyseisessä puussa kaikki $p < 0.001$, eli hyvin merkittäviä. Noodeissa ei näy muuttujien niiden nimiä, joten listaan ne tässä: Noodi 1 = Fi_cndmerk, Noodi 2 = Fi_TL_AM, Noodi 5 = Kääntäjä, Noodi 6 = Fi_TL_AM.

Kustakin noodista lähtevillä viivoilla on kerrottu, miten valitun muuttujan eri arvot on ositettu. Esimerkiksi Kääntäjä (Noodi 5) on jaettu kaikkiin *D-VT, IE, MT, PI, SA* sisältäviin arvoihin ja kaikkiin arvoihin, jotka sisältävät loput kääntäjät eli *OL, TE, UI, VL&IE, ZSHN*. Puun alaosassa näkyy pylväskuvioina osuudet *Yes* ja *No* arvoille näiden eri piirteiden yhdistelmien perusteella. Esimerkiksi *Noodi 3* näyttää, mitkä ovat konditionaalin käytön osuudet, kun valitsee ainoastaan ehtotulkintaisen *eht* konditionaalimerkityksen ja aikamuotona on perfekti. Suluissa oleva luku (n = 48) kertoo, montako tällaista havaintoa on.

Kuva 12. Ehdollisen päättelyn puu pylväskuvioina



Seuraavaksi kaksi esimerkkiä aineistoni konditionaaliesimerkeistä ja siitä, miten ne ovat jakautuneet puun eri noodeihin. Ensimmäinen esimerkki (esimerkki 14). Tässä esimerkissä on käännetty *tietää* verbin konditionaalin preesensmuoto (*cnd_pres*), joka on ehtotulkintainen (*eht*). Käännösratkaisuna on käytetty *verbiä*, joka on *indikatiivissa*. Eli ehdollisen päättelyn puun kuvassa (kuva 12) Noodissa 1 tämä kyseinen esimerkki kuuluu kategoriaan *No*. Se vaikuttaa Noodiin 2, koska siellä ovat kaikki konditionaaliesimerkit, jotka ovat ehtotulkintaisia. Ja se vaikuttaa vielä Noodin 4 konditionaalijakaumaan, koska siellä ovat kaikki ehtotulkintaiset (**eht**) preesensissä (**cnd_pres**) olevat konditionaaliesimerkit.

(14) …Gerda joka vain tuhahti, kun Juudit kuiskutteli olevansa huolissaan siitä, mitä hänen aviomiehensä mahtaisi ajatella, jos **_tietäisi_** vaimonsa kujertelevan julkisesti vieraan miehen kainalossa… (S. Oksanen, Kun kyyhkyset katosivat) / Герды, которая только фыркнула, когда Юдит поделилась с ней своей тревогой: а вдруг ее законный муж **узнает**, что она открыто гуляет с другим мужчиной? (Kogda isčezli golubi, kääänt. A. Sidorova,)

Toinen esimerkki samasta jakautumisesta (15). Tässä esimerkissä on käännetty *antaa* verbin konditionaalin preesensmuoto (*cnd_pres*), joka on todenvastaisuutta ilmaiseva (*tvast*). Käännösratkaisuna on käytetty *konditionaalia*, joka on muodostettu dat′ (дать) -verbin perfektimuodosta (***-l- muodosta)*** ja *by* partikkelista. Ehdollisen päättelyn puun kuvassa (kuva 12) Noodissa 1 tämä esimerkki kuuluu kategoriaan *Yes*. Se vaikuttaa Noodiin 5, koska siellä ovat kaikki todenvastaisuutta ilmaisevat konditionaaliesimerkit. Kääntäjänä on OL eli kyseinen esimerkki vaikuttaa myös Noodiin 6, koska siellä ovat kaikki kääntäjän OL kääntämät konditionaaliesimerkit. Ja lisäksi kyseinen esimerkki vaikuttaa vielä Noodi 8 konditionaalijakaumaan, koska siellä ovat kaikki todenvastaisuutta ilmaisevat (**tvast**), kääntäjän **OL** kääntämät ja konditionaalin preesensissä (**cnd_pres**) olevat konditionaaliesimerkit.

(15) Väätäinen vilkaisi minua. Katseesta näin, että se **_antaisi_**. Julkisuudesta on monenlaista hyötyä. (A. Salminen, Ei-kuori) / Вяятяйнен взглянула на меня. По взгляду выло видно, что она **бы дала**. В известности много плюсов. (Spasibo, net, käänt. L. Olykajnen)

Näiden kahden esimerkin tavoin ehdollisen päättelyn puun algoritmi on jakanut kaikki 1000 aineistoni konditionaaliesimerkkiä selittävien muuttujien (*Kääntäjä*, konditionaalimerkitys *Fi_cndmerk* ja konditionaalin aikamuoto *Fi_TL_AM*) mukaan.

Ehdollisen päättelyn puulle voi myös tehdä ennusteen (kuva 13). Tämä tehdään *predict* funktiolla, josta saadaan ennuste. Ristiintaulukoimalla ennusteen arvot ja havainnoidut arvot, voi nähdä kuinka hyvin saatu puu edustaa aineistoa.

Kuva 13. Puun ennustetaulukko

```
     pred.tree
       No Yes
No    705   9
Yes   247  39
```

Ennustetaulukossa *pred.tree* on nimi, jonka annoin taulukolle. Tämä taulukko (taulukko 14) selittää puun ennustetaulukon (kuva 13) ja virhematriisin (kuva 14) ristiintaulukoinnin tulokset:

Taulukko 14. Ennustetaulukon & virhematriisin (kuva 16) selitetaulukko

	Ennustettu: No	Ennustettu: Yes
Todellinen: No	ON = 705	VP = 9
Todellinen: Yes	VN = 247	OP = 39

Oikea positiivinen (OP) ennuste eli tapaukset, joissa ennustettiin, että konditionaalia käytetään ja joissa sitä myös käytettiin. Oikea negatiivinen (ON) ennuste eli tapaukset, joissa ennustettiin, että konditionaalia ei käytetä ja joissa sitä ei käytetty. Väärä positiivinen (VP) ennuste eli tapaukset, joissa ennustettiin, että konditionaalia käytetään, mutta sitä ei käytetty. Väärä negatiivinen (VN) ennuste eli tapaukset, joissa ennustettiin, että konditionaalia ei käytetä, mutta sitä käytettiin. (Banerjee yms. 2009, 130.)

Levshinan (2015, 297) mukaan voidaan laskea oikeiden ennusteiden prosenttimäärä, jos taulukon tuloksista oikea negatiivinen ja oikea positiivinen lasketaan yhteen, ja jaetaan eri tapausten kokonaismäärällä. Tässä siis lasketaan: (705 + 39) / 1000, josta vastaukseksi saadaan 0,744 eli 74,4 prosenttia. Eli oikeat ennusteet tehdään tämän puun tapauksessa 74 prosentille 1000 havainnosta.

Mielenkiintoista on huomata, että ehdollisen päättelyn puun algoritmi jakaa myös muuttujan Kääntäjä kahtia niissä tapauksissa, joissa konditionaalimerkitys on *tvast, mahd, tht*. Tämä johtuu siitä, että kuten Noodi 9 näyttää, niin kääntäjät *OL, TE, UI, VL&IE* ja *ZSHN* käänsivät, aikamuodosta riippumatta, yli 40 prosenttia ajasta konditionaalilla. Noin tuplasti enemmän kuin kääntäjien *D-VT, IE, MT, PI* ja *SA* konditionaalin käytön keskiarvo (n. 20 %) noodissa 6.

Ehdollisen päättelyn puu ei oikeastaan anna uutta informaatiota, mutta tässä sama tieto on esitetty selkeästi ja aineistossa olevat eri kaavat ovat helposti havainnoitavissa. Puuta tarkastellessa huomaa, että aikaisemmat havaintoni siitä, että ehtotulkintainen konditionaali kääntyy huomattavasti useammin konditionaalilla, kuin muun konditionaalimerkityksen omaavat konditionaaliesimerkit. Noodi 3 pylväskuviossa (kuva 12) näkyy lisäksi aiemmin mainitsemani asia eli se, että kun konditionaalimerkitys on ehtotulkintainen ja aikamuoto on perfekti, niin yli 80 prosenttia tapauksista käännetään konditionaalilla.

Myös suomenkielisen konditionaalimuodon perfektinen aikamuoto vaikuttaa aiemmin havaitsemallani tavalla konditionaalin käyttöön käännöksissä (käytetään enemmän). Jos vertaa noodeja 3 ja 4 sekä noodeja 7 ja 8, niin huomaa, että aikamuodon ollessa perfekti (3 & 7), on konditionaalin käyttö käännöksissä noin kaksi kertaa todennäköisempää, verrattuna preesensin noodeihin (4 & 8). Tämä näyttää lisäksi sen, että konditionaalimerkityksellä ei ole vaikutusta tähän jakaumaan.

4.5.2 Satunnaisen metsän analysointi

Loin satunnaisen metsän R:ssä käyttäen funktiota *randomForest()*, joka kuuluu saman nimiseen kirjastoon eli *randomForest*. Käytin siinä jälleen samoja muuttujia eli vastemuuttujaa Cond_Ru ja selittäviä muuttujia Kääntäjä, Fi_cndmerk, Fi_TL_AM. Loin metsän 1500 puusta, jossa kaikissa osituskohdissa testattiin kaikkia kolmea eri selittävää muuttujaa.

Metsän OOB (out-of-bag) virhemäärän arvio on 25,7 %. Jokainen metsän puu luodaan erillisestä näytteestä, joka otetaan aineistosta. Otettu näyte sisältää aina noin kaksi kolmasosaa 2/3 aineistosta. OOB eli out-of-bag virhearvio tehdään sen aineiston osuuden perusteella, joka jätetään kutakin metsän puuta luodessa pois eli noin kolmasosa aineistosta 1/3 (kaksi kolmasosaa ovat ns. in the bag). Arvioidun virhemäärän saa, jos laskee yhteen väärän negatiivisen (205) ja väärän positiivisen (52) ja jakaa summan kokonaismäärällä: (205 + 52) / 1000 = 0,257 (25,7 %).

Virhe- tai confusion-matriisin (confusion matrix) (kuva 14) luvut kertovat sen, kuinka monta mallin tapausta ennustettiin oikein. Se koostuu samoista asioista kuin yksittäisen puun ennustetaulukossa (kuva 14) näkyvä ristiintaulukointi. Eli tässä ovat ristiintaulukoituna metsän havainnoidut todelliset arvot ja ennustetut arvot. Satunnaisen metsän ennusteiden tarkkuuden voi laskea samalla tavalla kuin yksittäisen puun tapauksessa: (662 + 81) / 1000 = 0,743 eli 74,3 %. Tämä vastaa lähes täydellisesti yksittäisestä puusta saamaani oikeiden ennusteiden määrää 74,4 %. Satunnaisen metsän tuloksessa näkyy metsän puiden yhteen sulautettu (keskiarvoinen) tulos.

Kuva 14. Satunnainen metsä Cond_RU

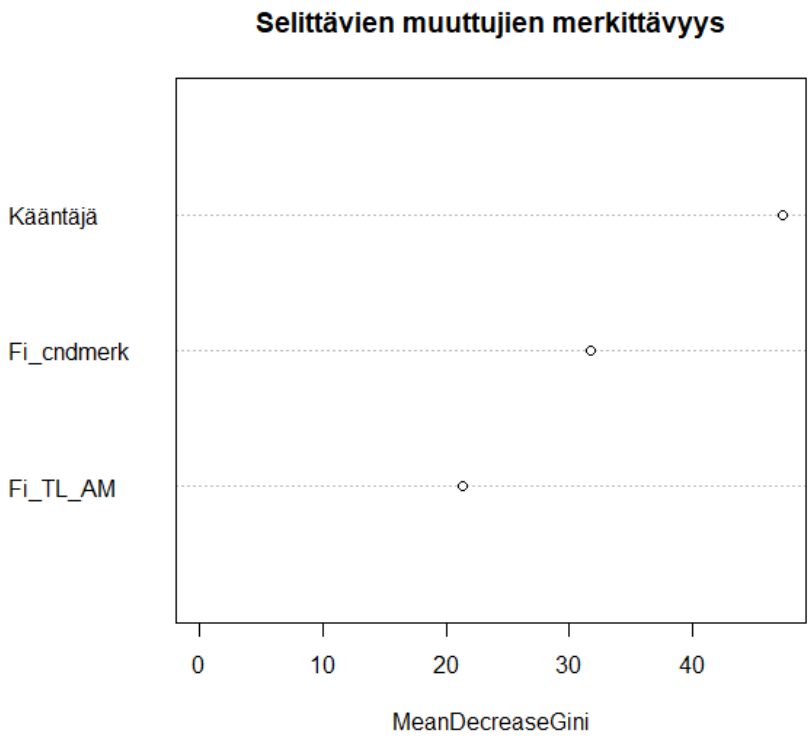
```
                Type of random forest: classification
                      Number of trees: 1500
No. of variables tried at each split: 3

        OOB estimate of  error rate: 25.7%
Confusion matrix:
     No Yes class.error
No  662  52  0.07282913
Yes 205  81  0.71678322
```

Luomani satunnaisen metsän puissa Cond_RU muuttujan arvot jakautuvat keskimäärin 1000 tapauksen kesken siis seuraavasti: 662 tapauksessa ennustettiin, että konditionaali ei käytetä ja sitä ei käytetty (ON); 205 tapauksessa ennustettiin, että konditionaalia ei käytetä, mutta todellisuudessa sitä käytettiin (VN); 52 tapauksessa ennustettiin, että konditionaalia käytetään, mutta sitä ei käytetty (VP); 81 tapauksessa ennustettiin, että konditionaalia käytetään ja sitä myös käytettiin (OP).

Tarkastellaan seuraavaksi metsän puiden muuttujien merkittävyyttä:

Kuva 15. Metsän puiden selittävien muuttujien merkittävyyden keskiarvo



Tärkein selittävä muuttuja metsäni puissa on siis Kääntäjä, joka saa merkittävyyden arvon 46,934. Tätä seuraa muuttuja Fi_cndmerk merkittävyysarvolla 31.798 ja pienimmän merkitysarvon 21.644 saa muuttuja Fi_TL_AM. Nämä arvot kertovat jokaisen selittävän muuttujan vaikutuksesta vastemuuttujaan muiden selittävien muuttujien yhteydessä. Metsän tulos on mielenkiintoinen, koska ehdollisen päättelyn puussa (kuva 12) on ylimpänä (tärkeimpänä) on Fi_cndmerk. Luomani satunnaisen metsän perusteella tärkein selittävä muuttuja ei kuitenkaan ole konditionaalimerkitys. Metsän 1500 puun perusteella konditionaalilla kääntämiseen vaikuttaa eniten kääntäjä. Tämä on tavallaan odotettu tulos siksi, koska mikään muu ei voi vaikuttaa enemmän konditionaalilla kääntämiseen kuin kääntäjä. Hän on se, joka tekee käännösratkaisut.

Tärkein vaikuttaja konditionaalin käyttöön käännöksessä on metsän perusteella siis *Kääntäjä*, joka vaikuttaa lähes puolessa tapauksista (46,9 %). *Fi_cndmerk* vaikuttaa noin kolmasosassa (31,7 %) tapauksista ja konditionaalin aikamuoto *FI_TL_AM* vain 21,6 prosenttia ajasta.

4.6 Vahvistuiko hypoteesi?

Oletin hypoteesissani, että suomen kielen konditionaalimuoto on käännetty suurimmaksi osaksi muilla tavoin, kuin venäjän kielen konditionaalilla. Tämä pitää aineistoni perusteella paikkansa. Suomenkielisistä konditionaaliesimerkeistä käännettiin venäjän kielen konditionaalilla ainoastaan 28,6 prosenttia, eli 71,4 prosenttia käännettiin muilla tavoin.

En osannut hypoteesissani sanoa, mikä on yleisin tapa kääntää suomen kielen konditionaali venäjän kielelle, mutta en uskonut sen olevan venäjän kielen konditionaalimuoto. Aineistoni tilastollisen analyysin perusteella oletukseni oli oikea. Yleisin tapa kääntää konditionaali venäjään on seuraava: *aikamuodoltaan futuurissa oleva perfektiaspektinen verbi, jonka tapaluokkana on indikatiivi.* Tällä tavoin on käännetty 24,1 prosenttia esimerkeistä.

Hypoteesissani oletin myös, että perfektiaikamuotoinen suomen kielen konditionaali kääntyy useammin venäjän kieleen konditionaalilla, kuin preesensaikamuotoinen konditionaali. Tämänkin osalta hypoteesini oli oikea. Perfektiaikamuotoisista konditionaaliesimerkeistä 44 % käännettiin venäjän kieleen konditionaalilla, kun taas preesensaikamuotoisista esimerkeistä ainoastaan 24 % käännettiin konditionaalilla. Hypoteesissani en osannut olettaa, että konditionaalimerkityksellä olisi niin suurta merkitystä, kuin sillä on konditionaalin käyttöön käännösvastineena.

5. PÄÄTELMÄT

5.1 Tutkimuksen yhteenveto

Aineisto koostuu 1000 konditionaaliesimerkistä, jotka haettiin korpuksesta satunnaisessa järjestyksessä. Jotkut kirjailijat ja luonnollisesti heidän teostensa kääntäjät ovat silti aineistossa yliedustettuja. Aineiston esimerkkien määrä ja kirjailijoiden teosten sanamäärä ParFin osakorpuksessa vastaavat kuitenkin hyvin toisiaan, joten näytteen pitäisi olla hyvä.

Aineiston käännöksistä 88,7 % tehtiin verbillä. Käännösvastineista 28,6 % on käännetty konditionaalilla ja 71,4 % muilla tavoin.

Logistisen regressiomallin tavoitteena oli selvittää, mitkä tekijät vaikuttavat siihen, että venäjänkielinen vastine on konditionaalissa. Optimaaliseksi valitsemani regressiomalli sisälsi kolme selittävää muuttujaa: Kääntäjä, Fi_cndmerk (konditionaalimerkitys), Fi_TL_AM (aikamuoto), jotka kaikki vaikuttavat vastemuuttujaan Cond_RU (konditionaalilla kääntäminen) jollain tavalla.

Kääntäjä -muuttujan vaikutus: Binäärisen logistisen regression tulosten perusteella kääntäjä henkilönä vaikuttaa konditionaalin esiintymiseen ja tietyt kääntäjät saattavat suosia jotain tiettyä käännösratkaisua. Aineistoni perusteella voi kuitenkin päätellä, että keskiarvo konditionaalin käytölle käännösratkaisuna on, kääntäjästä riippumatta, noin 25 %. Eli noin neljäsosa konditionaalin esiintymistä suomenkielisissä kaunokirjallisissa teksteissä käännetään venäjän kieleen konditionaalilla. Aineistoni konditionaalin käytön jakauma vastaa tukee tätä: 28,6 % käännetty konditionaalilla.

Fi_cndmerk -muuttujan vaikutus: Aineistoni konditionaalimerkityksistä intensionaalista asiaintilaa ilmaiseva *mahd,* todenvastaista tai toteutumatonta asiantilaa ilmaiseva *tvast* ja tahtotulkintainen *tht* vaikuttavat konditionaalin esiintymiseen käännöksissä negatiivisesti eli vähentävät sen käyttöä. Ehtotulkintainen konditionaalimerkitys *eht* vaikuttaa konditionaalin käyttöön positiivisesti eli lisää konditionaalin käyttöä käännösratkaisuna.

Fi_TL_AM -muuttujan vaikutus: Suomenkielisen konditionaalimuodon aikamuoto vaikuttaa vahvasti siihen käytetäänkö käännösratkaisuna konditionaalia vai ei. Jos konditionaalimuoto on aikamuodoltaan perfektissä, on aineistoni perusteella lähes 2 kertaa todennäköisempää, että käännösratkaisuna käytetään konditionaalia.

Näillä kolmella selittävällä muuttujalla on myös vaikutus toisiinsa konditionaalin käytön yhteydessä.

Konditionaalimerkityksellä on vaikutusta siihen käyttääkö kääntäjä käännösratkaisuna konditionaalia. Ehtotulkintainen konditionaalimerkitys vaikuttaa kaikkien kääntäjien konditionaalin käyttöön positiivisesti (52 % kaikista kääntäjien kääntämistä ehtotulkintaisista konditionaaliesimerkeistä on konditionaalilla käännettyjä) ja muiden merkitysten osalta konditionaalin käytön keskiarvo on 26,35 % (mahd = 23,8 %, tvast = 24 %, tht = 31,25 %). Ehtotulkintaisen konditionaalimerkityksen kääntyminen useimmiten konditionaalilla ei ole tutkimuksen perusteella siis kääntäjästä riippuva asia.

Aikamuodolla on, konditionaalimerkityksestä riippumatta, vaikutus konditionaalin käyttöön käännösratkaisuna. Kaikki konditionaalimerkitykset saavat korkeamman konditionaalin käytön todennäköisyyden ennusteen, kun aikamuoto on perfekti. Ehtotulkintaisen konditionaalin ollessa aikamuodoltaan perfektissä, on ennuste konditionaalin käytölle käännösratkaisuna yli 80 %.

Satunnaisen metsän tulosten perusteella kääntäjä vaikuttaa konditionaalin käyttöön käännösratkaisuna lähes puolessa (46,9 %) tapauksista. Konditionaalimerkitys vaikuttaa lähes kolmasosassa (31,7 %) ja konditionaalin aikamuoto noin viidesosassa (21,6 %) tapauksista.

5.2 Milloin konditionaalia tarvitaan venäjään kääntäessä

Venäjän kielen konditionaalilla on käännetty vähintään yksi kappale lähes kaikkia eri yhdistelmiä (aikamuoto, konditionaalimerkitys ja verbin semanttinen merkitys), jotka muodostavat suomen kielen konditionaalin aineistoni esimerkeissä. On muutamia, joita ei aineistossani esiinny, joista yhtenä esimerkkinä: semanttiselta merkitykseltään havaintoa ilmaisevasta verbistä rakentuva konditionaalin perfektissä oleva intensionaalista asiantilaa ilmaiseva konditionaalimuoto. Uskon kuitenkin, että aineistoni ulkopuolisista käännöksistä löytyy konditionaalilla käännettynä tämmöisiäkin yhdistelmiä, ainakin yksittäisiä esimerkkejä.

Yksittäisiä esimerkkejä löytyy toisaalta varmaan melkein mistä tahansa asiasta maailmassa. Joten totean mieluummin, että tutkimukseni perusteella konditionaalia tarvitaan eniten venäjään kääntäessä silloin, kun käännetään suomen kielen konditionaalia, joka on perfektissä, konditionaalimerkitykseltään ehtoa ilmaiseva eli ehtotulkintainen ja muodostuu olemista tai tekemistä ilmaisevasta verbistä.

5.3 Pohdintaa

Olen aina kuvitellut, että kielten välillä kielimuotojen (esim. konditionaali) kääntämisessä käytettäisiin yleisimmin kohdekielen vastaavaa kielimuotoa. Onhan se loogista ajatella näin. Voin

kuitenkin tutkimukseni perusteella todeta sen tosiasian, että näin ei aina ole. Käännökset ovat loppujen lopuksi aina ihmisten tekemiä ja samalla tavalla kuin me ihmiset emme ole kaikki samanlaisia, niin eivät ole myöskään meidän tekemämme käännösratkaisut. Sama pätee siis myös konditionaalin käyttöön käännösratkaisuna. Jotkut kääntäjät suosivat sitä, toiset eivät.

Tutkimukseni tuloksia voidaan mielestäni käyttää ainakin todisteena, kun halutaan väittää jotain konditionaalin käytöstä käännösratkaisuna venäjään kääntäessä. Toisaalta tutkimukseni tuloksilla on omat ongelmansa. Esimerkiksi konditionaalimerkitys sattuu olemaan yksi tärkeimmistä vaikuttajista konditionaalin käyttöön käännösratkaisuna. Olen jakanut aineistoni konditionaaliesimerkit konditionaalimerkityksen mukaan täysin yksin, eikä kukaan muu ole tarkistanut tässä luokittelussa tekemiäni ratkaisuja. Voi siis olla, että saamani tulokset ovat jossain määrin virheellisiä omien luokitteluvirheideni takia.

Tutkimukseeni liittyvänä mahdollisena jatkotutkimusaiheena olisi mielenkiintoista ottaa aiheeksi venäjän kielen konditionaalin kääntäminen suomen kielelle. Käytettäisiinkö siinä käännösratkaisuna useimmin suomen kielen konditionaalia vai mahdollisesti muita käännösratkaisuja ja mitä ne olisivat. Yhtenä jatkotutkimuksena olisi mielenkiintoista selvittää, vaikuttaako kääntäjän tausta käännösratkaisuihin, esim. koulutus, työkokemus, ikä yms. Tai selvittää, että käytetäänkö konditionaalia käännösratkaisuna enemmän, jossain toisessa tekstilajissa. Voisi olla myös mielenkiintoista tutkia, että kuinka paljon eri venäjän kielen konditionaalin muotoja käytetään käännöksissä *(-l-* muoto + *by*, predikatiivi + *by,* yms.).

Olisi mahtavaa, jos voisi tehdä jatkotutkimuksen niin, että kaikki eri konditionaaliesimerkkien ja käännösvastineiden luokittelut tarkistaisivat useampi henkilö. Tutkimuksen voisi jopa tehdä tällä samalla aineistolla, mutta mieluummin isommalla samankaltaisella aineistolla (esim. 5000 konditionaaliesimerkkiä). Tässä tapauksessa kaikki tutkimusta tekevät henkilöt voisivat tarkistaa toistensa luokittelumerkinnät ja pohtia niitä yhdessä. Isompi aineisto antaisi lisäksi vielä tarkempaa tietoa konditionaalin käytöstä käännösratkaisuna.

LÄHTEET

Tutkimusaineisto

ParFin korpus. Informaatioteknologian ja viestinnän tiedenkunta. Kielten yksikkö, Tampereen yliopisto.
< https://mustikka.uta.fi/texthammer> [viitattu 25.9.2018]

Tieteelliset lähteet

Alsohybe, Nabeel T. & Dahan, Neama Abdulaziz & Ba-Alwi, Fadl Mutaher 2017. *Machine-Translation History and Evolution: Survey for Arabic-English Translations*. Current Journal of Applied Science and Technology. 23:4, 1-19.

Banerjee, Amitav & Chitnis, U. B. & Jadhav, S. L. & Bhawalkar, J. S. & Chaudhury S. 2009. *Hypothesis testing, type I and type II errors*. Industrial Psychiatry Journal. 18:2, 127-131.

Cheng, Winnie 2012. *Exploring corpus linguistics: language in action*. London: Routledge.

Giamperi, Patrizia 2018. *Online Parallel and Comparable Corpora for Legal Translations*. Altre Modernità. 20, 237-252.

Henisz-Dostert, Bozena & Macdonald, R. Ross & Zarechnak, Michael 1979. *Machine Translation*. Mouton.

Hoey, M. & Houghton, D. 1998. Contrastive analysis and translation. Teoksessa M. Baker (Ed. assisted by K. Malmkjær) (toim.), *Routledge Encyclopedia of Translation Studies.* London & New York: Routledge. 45–49.

Hosmer, David W. & Lemenshow, Stanley 2004. *Applied Logistic Regression*. John Wiley & Sons, Incorporated.

Hutchins W. J. 1994. Machine Translation: History and General Principles. Teoksessa R. E. Asher & J. M. Y. Simpson (toim.), *The encyclopedia of languages and linguistics*. Oxford: Pergamon Press. 2322-2332.

Johansson, Stig 2007. *Seeing through Multilingual Corpora: On the use of corpora in contrastive studies*. Amsterdam: John Benjamins Publishing Company.

Kenny, Dorothy 1998. Theme and Rheme in Irish and English: A Corpus-based Study. Teoksessa M. Baker (Ed. assisted by K. Malmkjær) (toim.), *Routledge Encyclopedia of Translation Studies*. London & New York: Routledge. 50-53.

Kuusela, Kirsi 1992. *Soslagatel'noe naklonenie v sovremennom russkom jazyke i ego perevod na finskij jazyk : Konditionaali nykyvenäjässä ja sen kääntäminen suomen kielelle*. Pro gradu -tutkielma. Venäjän kieli ja kulttuuri. Tampere: Tampereen yliopisto.

Levshina, Natalia 2015. *How to do linguistics with R: data exploration and statistical analysis.* Amsterdam: John Benjamins Publishing Company.

McEnery, Tony & Hardie, Andrew 2012. *Corpus linguistics: method, theory and practice*. Cambridge University Press.

Merkle, Edgar & Shaffer, Victoria 2011. *Binary recursive partitioning: Backround, methods, and application to psychology*. British Journal of Mathematical & Statistical Psychology. 64:1, 161-181.

Mikhailov, Mikhail & Cooper, Robert 2016. *Corpus linguistics for translation and contrastive studies: A guide for research*. London: Routledge.

Moorkens, Joss 2013. *Consistency in Translation Memory Corpora: A Mixed Methods Case Study.* Journal of Mixed Methods Research 2015. 9:1, 31–50.

Pandis, Nikolaos 2016. *The chi-squared test.* American Journal of Orthodontics and Dentofacial Orthopedics. 150:5, 898-899.

Zaharov, Viktor & Bogdanova, Svetlana 2011. *Korpusnaâ lingvistika: učebnik dlâ studentov gumanitarnyh vuzov.* Irkutsk: IGLU.

Sanakirjat

Akademiâ nauk SSSR, institut russkogo âzyka 1980. *Russkaâ grammatika.* Moskova: Nauka,

Martin, Elizabeth A. & McFerran Tanya A. (toim.) 2017. *A Dictionary of Nursing.* Oxford: Oxford University Press. <https://www.oxfordreference.com/view/10.1093/acref/9780198788454.001.0001/acref-9780198788454> [viitattu 20.3.2019]

VISK: Iso suomen kielioppi verkkoversio. <http://scripta.kotus.fi/visk> [viitattu 15.10.2018]

Muut lähteet

De Sutter, Gert 2018. *Introduction to Advanced Corpus Analysis: Multivariate Statistics, Workshop Tools and Methods for Corpus-Based Translation Science.* Austria: University of Innsbruck <https://transbank.info/workshop/>

Dobrušina, N. R. 2014. *Soslagatel'noe naklonenie.* <rusgram.ru/pdf/3_subjunctive_dobrushina_20141102_finalfinal.pdf> [Luettu: 10.1.2019]

Finnlectura.fi. 2.5.2.1.4 Modukset <https://fl.finnlectura.fi/verkkosuomi/Morfologia/sivu25214.htm> [viitattu 15.10.2018]

KvantiMOTV: Kvantitatiivisten tutkimusmenetelmien menetelmäopetuksen tietovaranto <https://www.fsd.uta.fi/menetelmaopetus/mittaaminen/ominaisuudet.html> [viitattu 6.5.2019]

Shuttleworth, Martyn & Wilson, Lyndsay T. 2008. *Confounding Variable / Third Variable* <https://explorable.com/confounding-variables> [viitattu 6.5.2019]

Texthammer, ver. 1.5. User manual <https://mustikka.uta.fi/texthammer> [viitattu 20.9.2018]

Tilastokeskus <https://www.stat.fi/meta/kas/index.html> [viitattu 20.5.2019]

РЕФЕРАТ НА РУССКОМ ЯЗЫКЕ

Сослагательное наклонение в финском языке и его эквиваленты в финско-русских переводах

Корпусное исследование

1. ВВЕДЕНИЕ

1.1 Исследовательский вопрос

Само сослагательное наклонение исследовано очень подробно, как в финском, так и в русском языке, но его соответствия при переводе с финского на русский исследовались мало. Нам удалось найти лишь одну работу, по этой тематике. Это – дипломная работа Кирси Куусела (Kirsi Kuusela 1992) «*Сослагательное наклонение в современном русском языке и его перевод на финский*». Наше исследование посвящено переводу конструкций с сослагательным наклонением с финского на русский, то есть пара языков та же, но направление противоположное.

Первоначальной исследовательской проблемой данной работы стало как раз то, что таких исследований не существует. Таким образом, необходимо выяснить, какими разными способами финское сослагательное наклонение переводится на русский язык. По этим причинам, мы считаем, что это интересно и стоит исследовать, потому что эта тема мало изучена, и результаты исследования дадут новую информацию о переводе сослагательного наклонения и его использования в переводах.

Основные исследовательские вопросы нашей работы:

1. Насколько часто при переводах с финского на русский формы финского кондиционала переводятся формами русского сослагательного наклонения?
2. Какое из русских соответствий финского кондиционала используется больше всего при переводе?
3. Влияют ли грамматическое время, грамматическое или семантическое значение финского кондиционала на соответствия при переводе?

4. Влияют ли преференции переводчика на перевод финского кондиционала?

1.2 Материал исследования

Данное исследование сделано на основе корпусного материала. Фактический материал исследования состоит из 1000 примеров употребления сослагательного наклонения в финском языке и их переводов. Материал был получен из параллельного корпуса художественной литературы *ParFin,* собранного исследователями переводоведения Университета Тампере. Из корпуса ParFin был выбран подкорпус с текстами, опубликованными после 1976 года. Это ограничение было сделано для того, чтобы в материале не было примеров из очень старых текстов, таких как «Железная дорога» Юхани Ахо (Juhani Aho, «Rautatie» 1884). Затем из подкорпуса были получены 1000 случайных примеров использования сослагательного наклонения в финском языке и их переводы, т.н. параллельный конкорданс. (см. Kuva 1, 32)

Для квалитативного анализа в качестве первоначальной обработки материала в таблицу с параллельным конкордансом сослагательного наклонения были добавлены еще следующие столбцы (переменные/факторы): Переводчик (Kääntäjä), глагол финского кондиционала (Fi_form), грамматическое время финского кондиционала (Fi_TL_AM), грамматическое значение финского кондиционала (Fi_cndmerk), семантическое значение финского кондиционала (Fi_vclas), перевод на русский (Ru_form), часть речи перевода (Ru_SL), наклонение перевода (Ru_TL), грамматическое время перевода (Ru_AM), аспект перевода (Ru_ASP), лексическое средство, использованное для выражения сослагательного наклонения (Leks_ru).

Цель данного исследования – выяснить, какие факторы влияют на появление русского сослагательного наклонения в переводе. Таким образом в нашем исследовании бинарным предиктором является Cond_RU – эта переменная показывает, является ли перевод в сослагательном наклонении или нет. Факторами, возможно влияющими на предиктор, являются все остальные ранее перечисленные столбцы.

1.3 Методы

Основные методы исследования – *бинарная логистическая регрессия* (binary logistic regression) и *дерево условного вывода* (conditional inference tree) и *случайный лес* (random forest).

Бинарная логистическая регрессия предоставляет информацию о том, какие факторы (независимые переменные) оказывают существенное влияние на исследуемый предиктор (зависимая переменная), насколько сильно они влияют на предиктор, а также, является ли это влияние позитивным или негативным (см. De Sutter 2018).

Дерево условного вывода – это метод, основанный на бинарном рекурсивном разделении (Merkle & Shaffer 2011). Результатом анализа является граф в форме дерева, показывающий влияние различных факторов друг на друга и на предиктор.

Метод *случайного леса* заключается в генерировании из экспериментального массива данных путем незначительного изменения величин большого количества похожих массивов данных (например, 500) и создание для каждого из них дерева условного вывода, в результате чего возникает “случайный лес”. Результаты деревьев леса объединяются, и полученные значения показывают влияние разных факторов на предиктор в сочетании с другими факторами. Результатом является точное понимание того, какие факторы больше всего влияют на предиктор (см. Levshina 2015).

Фактическая обработка и анализ материала, с употреблением всех методов анализа, выполнялась с помощью программы «R» и ее графического интерфейса «RStudio». Для обработки были использованы инструкции, которые адъюнкт-профессор Герт Де Суттер (Gert De Sutter) дал в своей лекции об инструментах и методах для переводческого исследования, основанного на корпусных данных «Tools and Methods for Corpus-Based Translation Science». Для обработки материала в *RStudio* были использованы следующие библиотеки *R*: car, effects, lme4, MuMIn, MASS, party, partykit, rms, randomForest, e1071.

2. ТЕОРЕТИЧЕСКИЕ ОСНОВЫ ИССЛЕДОВАНИЯ

2.1 Корпусы и их использование в исследованиях

С появлением электронных текстов и доступа к колоссальным ресурсам Интернета появилось много возможностей для сбора текстов на разных языках для исследования различных языков.

«Под ***лингвистическим,*** или ***языковым, корпусом текстов*** понимается большой, представленный в машиночитаемом виде, унифицированный, структурированный, размеченный, филологически компетентный массив языковых данных, предназначенный для решения конкретных лингвистических задач.» (Zaharov & Bogdanova 2011, 7).

Существует много разных типов корпусов, и их можно классифицировать по разным критериям. Основываясь на *лингвистической информации*, корпусы можно разделить на текстовые, речевые и смешанные корпусы. Смешанные корпусы содержат, естественно, как текст, так и речь. Корпусы могут быть классифицированы на основе *параллельности*: одноязычные, двуязычные или многоязычные. По критериям *литературности* корпусы делятся на литературные, разговорные, диалектные, терминологические и смешанные корпусы. Корпусы могут также быть разделены по *назначению создания* корпуса: на многоцелевые корпусы и на корпусы для специальных исследовательских целей (Zaharov & Bogdanova 2011, 22-23).

Корпус ParFin (параллельный финско-русский корпус), использованный при изучении данного исследования, подразделяется на эти критерии следующим образом: ParFin это *текстовой корпус*, *двуязычный параллельный корпус*, по литературности *литературный корпус* (составлен из литературных произведений) и созданный для *различных целей*.

Полезность использования корпусов в исследованиях обусловлена следующими причинами. Достаточно большое представление (размер) корпуса обеспечивает достоверность получаемой из него информации и дает исчерпывающий обзор всех языковых явлений. Различная информация, найденная в корпусе, присутствует в ее истинном контексте, что дает возможность изучить их всесторонне и объективно. После сборки и подготовки корпус может много раз использоваться разными исследователями и для различных целей (Zaharov & Bogdanova 2011, 8).

Исследования с использованием многоязычных корпусов постоянно развиваются. Многоязычные корпусы используются, например, в следующих исследованиях:

контрастивная лингвистика, типология языка, переводоведение и подготовка переводчиков, двуязычная лексикография, преподавание иностранных языков и обработка естественного языка (включая машинный перевод) (Johansson 2007, 301).

2.2 Сослагательное наклонение

2.2.1 Сослагательное наклонение в финском языке

В финском языке сослагательное наклонение образуется от глагольной основы с помощью суффикса сослагательного наклонения *-isi-*. У него есть две временные формы: настоящее время и перфект (Finnlectura.fi 2.5.2.1.4.).

В сослагательном наклонении финского языка существует значение возможности, но оно отличается от других способов выражения модальности: описываемая ситуация не соответствует действительности, а зависит от воли и воображения говорящего, то есть является альтернативной. Другими словами, сослагательное наклонение отражает тот факт, что рассматриваемая ситуация является одной из альтернатив, а не реальностью (VIKS § 1590).

Сослагательное наклонение финского языка указывает на то, что вы делаете, как что-то условное или неопределенное. Оно может также выражать желание, убеждение, подозрение или вежливый запрос. В отличие от потенциального наклонения, сослагательное наклонение обычно присутствует в придаточных предложениях. Его использование широко распространено, особенно в условных выражениях (например, *Jos olisin kuningas, asuisin linnassa. - Если бы я был королем, я бы жил в замке.*) Условным предложением выражается условие, при котором, утверждение, которое в настоящее время не соответствует действительности, является истиной (Finnlectura.fi 2.5.2.1.4.).

Финское сослагательное наклонение используется для выражения разных ситуаций. Им можно выразить запланированную ситуацию (**mahd**). В таких случаях сослагательное наклонения выражает, что запланированная ситуация или положение вещей является одним из возможных вариантов (например, *Näistä saappaista saisi hyvät kalossit. – Из этих сапог можно было бы сделать хорошие калоши.*) (VISK § 1592).

Сослагательное наклонение также используется при выражении просьбы или предложения (**tht**). Это значение возникает в ситуациях, которые выражают просьбу, запрос, предложение или приглашение (например, *Kävisit nyt siellä jo. – Сходил бы ты уже туда.*) (VISK § 1593).

У финского кондиционала существует интерпретация ложной и нереальной модальности (**tvast**). Таким образом формы настоящего времени и перфекта кондиционала интерпретируются, например, в риторических вопросах и в придаточных предложениях отрицательных вопросов (например, *En muista, milloin olisin nukkunut näin sikeästi. – Я не помню, когда бы я спал так сладко*.) Кроме того, перфект кондиционала указывает упущенную возможность в прошлом (например, *Olisit heittänyt pidemmälle. – Бросил бы ты дальше.*) и еще нереализованный, например, запланированный вариант (например, *Korjailisin mieluummin pyörää tässä vähän. - Да я лучше немного велик починил бы.*) (VISK § 1593).

В финском языке сослагательное наклонение используется еще и в условных предложениях (**eht**). Условные предложения нужны, когда речь идет о вещах, которые, так или иначе, зависят друг от друга. В предложениях, выражающих эту условность, необязательно присутствует частица *если* (jos), например, *Kuulisit tuon kappaleen niin tietäisit, miksi pidän siitä. (jos kuulisit) – Услышал бы ты эту песню, то ты понял бы, почему она мне нравится. (если услышал*) (VISK § 1595).

2.2.2 Сослагательное наклонение в русском языке

В русском языке сослагательное наклонение выражается частицей *бы (б).* Оно включает в себя различные комбинации (союзы) частицы *бы (б)*, такие как *чтобы (б)* и устаревшие *кабы* и *дабы.* То есть, образующая сослагательное наклонение частица *бы* может появляться в полной или сокращенной форме *б,* или в составе союза *чтобы.* Полная форма *бы* не имеет морфологических ограничений для комбинации, но сокращенная форма *б* имеет формальные ограничения на контекст употребления: *б* нельзя использовать после слова, заканчивающегося согласной (Dobrušina 2014, 68).

Сослагательное наклонения русского языка не имеет форм времени (Russkaâ grammatika 1980, 625). Таким образом, ситуация, отраженная сослагательным наклонением, может одновременно отражать настоящее, прошедшее и будущее время.

Частица *бы (б)* и союз *чтобы* могут употребляться с формами прошедшего времени глагола, с инфинитивами, с предикативами, в составе эллиптических конструкций, с причастиями, с деепричастиями и с императивами (Dobrušina 2014, 68).

Наиболее распространенными являются комбинации прошедшего времени глагола и частицы *бы (б)* (Dobrušina 2014, 68). Это часто называется **формой на -л**, потому что она совпадает по форме с прошедшим временем индикатива, например *прыгал бы* (Russkaâ grammatika 1980,

625). Таким образом, можно отличить форму прошедшего времени глагола (**форма на -л**), которая образует сослагательное наклонение, от индикативной формы прошедшего времени глагола, поскольку сослагательное наклонение не имеет форм времени. Однако в описании сослагательного наклонения принято называть эту **форму на -*л*** формой прошедшего времени глагола (Dobrušina 2014, 68).

По этим причинам и в данной работе, когда речь идет о форме **прошедшего времени глагола**, то имеется ввиду именно эта **форма на -л.**

Сослагательное наклонение отражает ситуации, которых нет в реальном мире. Эта ситуация может быть по значению контрфактивной (например, *На вашем месте я бы этого не делал.*) или выражать желательность (например, *Только бы он не услышал.*) (Dobrušina 2014, 66).

Сослагательное наклонение может также иметь прагматические функции в переносных использованиях: рассказать о намерениях говорящего смягченно (например, *Я бы попросил вас не кому не рассказывать об этом*), и при попытке снизить категоричность аргумента (*Я бы сказал, что это воровство*). Сослагательное наклонения часто используется в придаточных предложениях (Dobrušina 2014, 66).

3. РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ

Из 1000 примеров материала 777 (77,7 %) употреблены в форме кондиционала настоящего времени и 223 (22,3 %) – в форме кондиционала перфекта. В 887 примерах (88,7 %) финские кондиционалы были переведены на русский с помощью форм глагола. И 286 (28,6 %) примера были переведены сослагательным наклонением, а 714 (71,4 %) – другими способами.

При переводе, из русских соответствий финского кондиционала, чаще всего используется **глагол совершенного вида** в **будущем времени индикатива**. Таких переводов в нашем материале 24,1 процента.

3.1 Результаты бинарной логистической регрессии

С помощью бинарной логистической регрессии нам удалось установить факторы, которые значительно влияют на наш предиктор, то есть на использование в переводе сослагательного наклонения *Cond_RU*. Значительно влияющими факторами оказались: Переводчик (Kääntäjä), грамматическое значение финского кондиционала (Fi_cndmerk) и грамматическое время финского кондиционала (Fi_TL_AM).

3.1.1 Влияние фактора *Kääntäjä* на предиктор *Cond_RU*

Из 1000 примеров финского кондиционала в нашем материале, 805 примеров перевели 4 переводчика из 10. Употребление сослагательного наклонения в переводе у этих четырех переводчиков в среднем составляет всего 24,43 процента, то есть 75,57 процента они переводили другими способами.

На основании этого можно сделать выводы о влиянии фактора *Переводчик* на употребление сослагательного наклонения в переводах. Конечно, переводчик лично влияет на употребление сослагательного наклонения в переводах, так как именно он принимает решения для переводов различных языковых форм, и может предпочитать какие-то из этих решений. На основании данного материала можно сказать, например, что переводческий дует Зайкова и Хотинского (ZS&HN) предпочитает переводить кондиционал финского языка русским сослагательным наклонением. Они сделали 56,52 процента из переводов этим способом.

Тем не менее, мы предполагаем, что по мере увеличения числа примеров кондиционала, употребление сослагательного наклонения в переводах будет выравниваться независимо от переводчика, и в итоге получится распределение, где 25 процентов переведено сослагательным наклонением, а 75 процентов другими решениями для перевода. Это также

подтверждается употреблением сослагательного наклонения в переводах данного материала, поскольку из примеров кондиционала в нашем материале 28,6 процента были переведены на русский язык сослагательным наклонением.

3.1.2 Влияние фактора *Fi_cndmerk* на предиктор *Cond_RU*

В таблице результатов моей оптимальной модели бинарной логистической регрессии грамматические значения кондиционала, выражающие просьбу (tht), запланированную ситуацию (mahd) и ложную или нереальную модальность (tvast), влияют негативно на предиктор *Cond_RU* (употребление сослагательного наклонения в переводе). Позитивно на предиктор влияет только условное значение кондиционала (eht).

Другими словами, основываясь на нашем материале, мы можем утверждать, что грамматическое значение кондиционала влияет на употребление сослагательного наклонения в переводе на русский язык. Но из значений кондиционала только условное значение переводится на русский язык чаще сослагательным наклонением (52,3 %), чем другими способами. Для остальных значений эффект обратный, то есть они переводятся сослагательным наклонением более редко (в среднем только 26,35 % были переведены сослагательным наклонением).

3.1.3 Влияние фактора *Fi_TL_AM* на предиктор *Cond_RU*

Из перфектных примеров финского кондиционала (223), 99 примера (44 %) были переведены на русский язык сослагательным наклонением. А из примеров финского кондиционала, которые в настоящем времени (777), всего лишь, 187 примера (24 %) были переведены на русский язык сослагательным наклонением.

На основании этой информации, можно утверждать, что грамматическое время финского кондиционала сильно влияет на перевод. Поскольку, например, при перфекте финского кондиционала в нашем материале, почти в 2 раза более вероятно, что перевод будет сделан сослагательным наклонением.

3.2 Результаты *дерева условного вывода* (conditional inference tree)

Дерево условного вывода (см. Kuva 12) на самом деле не дает новой информации, но в нем та же информация представлена в виде графа.

Глядя на дерево, можно заметить наши ранние наблюдения о том, что условное значение финского кондиционала переводится на русский язык гораздо чаще, чем другие значения. Таким же образом, можно увидеть, что при переводе финского перфектного кондиционала сослагательное наклонение используется больше. К тому же дерево показывает, что грамматическое значение финского кондиционала не влияет на это, так как при переводе перфектного кондиционала сослагательное наклонение употребляется больше, несмотря на то, какое грамматическое значение имеет финский кондиционал.

Смотря на дерево, мы видим так-же, что когда грамматическое значение кондиционала условное и время перфект, то более 80 процентов из примеров переводятся сослагательным наклонением.

3.3 Результаты *случайного леса* (random forest)

В данной работе случайный лес был сделан из 1500 деревьев условного вывода. Наиболее важным фактором случайного леса является *Kääntäjä* (Переводчик), степень значения которого составляет 46,934. Затем следует фактор *Fi_cndmerk* (грамматическое значение) со значимостью 31,798, а наименьшую значимость получает фактор *Fi_TL_AM* (грамматическое время) со значимостью 21,644.

Эти значимости указывают на влияние каждого фактора на предиктор, в сочетании с другими факторами. Результат леса интересен тем, что в отдельном дереве условного вывода самым значительным фактором является грамматическое значение кондиционала *Fi_cndmerk*. Однако, в случайном лесе самым важным фактором является переводчик *Kääntäjä*.

И так, наиболее важным фактором, влияющим на употребление сослагательного наклонения в переводе, является переводчик, который влияет почти в половине переводов (46,9 %). Грамматическое значение финского кондиционала влияет в 31,7 процентах переводов, а грамматическое время финского кондиционала влияет в 21,6 процентах переводов финского кондиционала.

4. ВЫВОДЫ

На основании результатов данного исследования можно сказать, что сослагательное наклонение наиболее необходимо при переводе финского кондиционала, который является по времени перфектным, по грамматическому значению условным и который образуется глаголом, выражающим бытие или действие.

В конце концов, переводы всегда делаются людьми, и, как люди, мы не одинаковы, так и наши решения при переводе разные. То же самое касается и употребления сослагательного наклонения в переводе. Некоторые переводчики предпочитают его, другие нет.

Результаты данной работы имеют свои проблемы. Например, грамматическое значение финского кондиционала оказывается одним из наиболее важных факторов, влияющих на использование сослагательного наклонения в качестве решения для перевода.

Разметка материала такого типа - сложная работа, требующая подготовки и тренировки. Поскольку, например, при семантической разметке обязательно есть маргинальные случаи, которые могут быть проинтерпретированы по-разному. В больших проектах могут использовать нескольких операторов и потом сравнивать их разметки между собой. В данном случае это невозможно, поэтому разметку выполнял один оператор, он же исследователь, он же - автор работы. Поэтому возможны неточности в разметке грамматического значения финского кондиционала.

В качестве возможных тем дальнейшего исследования, связанного с данной работой, было бы интересно исследовать перевод русского сослагательного наклонения на финский язык. Или изучить, что влияют ли данные переводчика (например, образование, стаж, возраст и т. п.) на употребление сослагательного наклонения в переводе. Или выяснить используется ли сослагательное наклонение в качестве решения для перевода более чаще в других типах текста. Также было бы интересно изучить, насколько часто разные формы русского сослагательного наклонения (например, форма на **-л**, предикатив и другие) используются в переводах.

Было бы очень интересно провести дополнительное исследование, таким образом, что все различные классификации примеров кондиционала и их переводов были бы проверены несколькими людьми. Это исследование можно было бы выполнить даже с материалом данной работы, но предпочтительно с похожим материалом большего объема (например, 5000 примеров). В таком случае все те, кто участвуют в проведении исследования, смогут проверять

классификации друг друга и рассматривать их вместе. Кроме того, материал большего объема предоставит еще наиболее точную информацию об употреблении сослагательного наклонения в качестве решения для перевода.